UNIVERSITY OF CALIFORNIA,
IRVINE

Data-Driven Models for Dynamics of Gene Expression and Single Cells

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Tao Peng

Dissertation Committee:
Professor Qing Nie, Chair
Professor Frederic Yui-Ming Wan
Professor Long Chen

2017

DEDICATION

To

my parents, my wife and my son

for your love, support and understanding

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

Many people have helped me during the past five years at UC Irvine. First, I would like to express my sincere appreciation and thanks to my advisor Professor Qing Nie for all insightful guidance and tremendous support during my time as a graduate student. I am so lucky to be his graduate student since he not only offered countless advice and showed the right directions when I got lost in the research but also encouraged me to explore my interests freely. During the time working with him I have learned how to look for the research questions, to search the means of solving the problems, to summarize the results, to present the work and to be an independent researcher. Moreover, his scientific attitude, personality, and work ethic have made a profound impact on me. It has been a real pleasure working with him, and I would not have made it through without his abundant help and confidence in me.

Second, many thanks to my collaborators Professor Weian Zhao, Professor Maksim Plikus, Professor Xiangmin Xu, Professor Xiaohui Xie, Professor Ali Mortazavi, Professor Xing Dai, Dr. Yulin Shi, Dr. Linan Liu, Dr. Ji Won Oh, Dr. Xiaojie Wang, Dr. Weihua Zeng, Dr. Ricardo Ramirez, Raul Ramos who always helped me better understand our biological problems. Without their help this work could not have been done. A special thanks to Professor Ali Mortazavi, Dr. Weihua Zeng, Dr. Ricardo Ramirez for helping me through the three month rotation in Professor Ali Mortazavi's lab. I learned so much knowledge about the next generation sequencing experiments and data analysis from Dr. Weihua Zeng and Dr. Ricardo Ramirez. They always answered my questions very patiently. And I would like to express my thanks to Prof. Weian Zhao and Dr. Linan Liu for answering my queries about our joint work and revising the manuscript. I would also express particular thanks to Professor Xiangmin Xu and Dr. Yulin Shi. Dr. Yulin Shi helped me a lot analyze the LSPS data using the matlab GUI package developed by him and I also enjoyed lots of discussion with him beyond the research. In particular, I would like to express my strong thanks to Professor Maksim Plikus, who gave me so many opportunities to involve the projects in his lab and from whom I have learned so much.

Third, I am also grateful to the other two members of my dissertation committee, Professor Frederic Wan and Professor Long Chen. I appreciate all the help they have provided for me during the past years and recommendation letters you wrote for me. Thanks to all the current and past members of the Nie lab for all their help. All of them are awesome people. Especially, many thanks to Dr. Chunhe Li for his help to me on understanding the landscape method which was employed in one of my projects. I would like to thank Catherine Ta, who provided lots of help to me, such as reviewing materials for the classes, taking us to the local Vietnamese restaurant, and so on. What's more, I am honored to thank Dr. Adam MacLean for revising the manuscript. In particular, I owe thanks for Dr. Tian Hong, Dr. Weitao Chen, and Dr. Huijing Du, who helped me tremendously go through the difficult time during the graduate school. It was fun having dinners and playing cards after dinners with them.

Forth, I am so appreciated to UC Irvine for its generous offer of scholarship in the past five

years. My dissertation would not exist without its full support. Specially thanks to Karen Martin, Naomi Carreon, Cely Dean from Center for complex Biological Systems and Donna McConnell from Department of Mathematics for their help.

Last but not least, my gratitude goes to my parents, my sisters, my wife and my son for their understanding and encouragement. I feel so loved and blessed and all of them are my rock.

I have the pleasure to express my thanks to everyone else who helped me along the way. I apologized for not being able to name each and every person who has had an impact on my life in the past five years.

My graduate research and this thesis have been supported in part by National Institute of Health grants P50GM76516, R01GM107264, and R01ED023050, and National Science Foundation grants DMS1161621 and DMS1562176.

CURRICULUM VITAE

Tao Peng

EDUCATION

Doctor of Philosophy in Mathematics **2017**
University of California, Irvine *Irvine, California*

Master of Science in Mathematics **2009**
Wuhan University *Wuhan, China*

Bachelor of Science in Mathematics **2007**
Wuhan University *Wuhan, China*

RESEARCH EXPERIENCE

Graduate Research Assistant **2013–2017**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2015–2016**
University California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Tao Peng*, Linan Liu*, Adam L MacLean, Chi Wut Wong, Weian Zhao, Qing Nie. A mathematical model of mechanotransduction reveals how mechanical memory regulates mesenchymal stem cell fate decisions. BMC Systems Biology. (*contributed equally to this work)

William R. Holmes, Nabora Soledad Reyes de Mochel, QixuanWang, Huijing Du, **Tao Peng**, Michael Chiang, Olivier Cinquin, Ken W.Y. Cho and Qing Nie. Gene expression noise enhances robust organization of the early mammalian blastocyst. PloS Computational Biology, 2017.

Xiangmin Xu, Nicholas D. Olivas, Taruna Ikrar, **Tao Peng**, Todd C. Holmes, Qing Nie, and Yulin Shi. Primary visual cortex shows laminar-specific and balanced circuit organization of excitatory and inhibitory synaptic connectivity. The Journal of Physiology, 2016.

MANUSCRIPT SUBMITTED

Tao Peng, Qing Nie. SOMSC: Self-Organization-Map for high-dimensional single-cell data of cellular states and their transitions. http://biorxiv.org/content/early/2017/04/06/124735

Qixuan Wang, Ji Won Oh, HyeLim Lee, Anukriti Dhar, **Tao Peng**, et al. A multi-scale model for the hair follicle reveals heterogeneous skin domains driving rapid spatiotemporal hair growth patterning. Submitted.

MANUSCRIPT IN PREPARATION

Tao Peng, Xing Dai, Qing Nie. Study controlling factors in mouse embryonic epidermal development.

Tao Peng, Qing Nie. Network inference incorporating the prior information.

ABSTRACT OF THE DISSERTATION

Data-Driven Models for Dynamics of Gene Expression and Single Cells

By

Tao Peng

Doctor of Philosophy in Mathematics

University of California, Irvine, 2017

Professor Qing Nie, Chair

This thesis uses mathematical models to study the dynamics of biological systems under the single cell level. In the first chapter we study a minimal gene regulatory network permissive of multi-lineage mesenchymal stem cell differentiation into four cell fates. We present a continuous model that is able to describe the cell fate transitions that occur during differentiation, and analyze its dynamics with tools from multistability, bifurcation, and cell fate landscape analysis, and via stochastic simulation. In the second chapter we adapt a classical self-organizing-map approach to single-cell gene expression data, such as those based on qPCR and RNA-seq. In this method, a cellular state map (CSM) is derived and employed to identify cellular states inherited in a population of measured single cells. Cells located in the same basin of the CSM are considered as in one cellular state while barriers between the basins provide information on transitions among the cellular states. Consequently, paths of cellular state transitions (e.g. differentiation) and a temporal ordering of the measured single cells are obtained. In the third chapter on the basis of the functional mapping assays of primary visual cortex, we conducted a quantitative assessment of both excitatory and inhibitory synaptic laminar connections to excitatory cells at single cell resolution, establishing precise layer-by-layer synaptic wiring diagrams of excitatory and inhibitory neurons in the visual cortex inferred by the mathematical model. In the fourth chapter we constructed a multi-scale mathematical model integrating the gene regulatory network and cell lineage to

study the functions of key genes in controlling mouse embryonic epidermis development. In the fifth chapter we studied the selections of models when prior information is provided to infer the gene regulatory network combining the expression data and ChIP-seq data.

Chapter 1

A mathematical model of mesenchymal stem cell fate decisions

[Chapeter 1 is reprinted with the permission from Tao Peng, Linan Liu, Adam L MacLean, Chi Wut Wong, Weian Zhao and Qing Nie. A mathematical model of mechanotransduction reveals how mechanical memory regulates mesenchymal stem cell fate decisions. BMC Systems Biology, 2017. ©2017 The Authors.[114]]

1.1 Background

Changes in cellular state can be regulated by mechanical signals from the cellular microenvironment, such as the local extracellular matrix (ECM) stiffness [35, 45, 148, 69]. Recent studies into mechanotransduction have demonstrated that cells sense and integrate mechanical cues from the ECM, causing transcriptional changes to occur and influencing cell fate decisions [35, 45, 148, 62]. Mesenchymal stem cells (MSCs) are controlled by signals from the ECM and exhibit a wide range of differential gene expression patterns [35, 49]. The mechanisms governing how MSCs sense the surrounding ECM, and the myriad other factors affecting MSC fate, including interactions with proteins and ligands, tethering, and porosity, remain incompletely defined [148, 160]. Further understanding of how differentiation cues are mediated by mechanical stimuli will help to facilitate new biomaterial design, cell-based therapeutics, and engineered tissue constructs for use in regenerative medicine.

The signals arising at the stem cell/substrate interface are complex and dynamic [160], however it has been shown that stiffness alone is enough to direct MSC differentiation [148, 69]. MSCs undergo neurogenic or adipogenic differentiation on soft substrates ($<$ 1 kPa), and myogenic or osteogenic differentiation on stiff substrates (>10 kPa) [35, 62] (Fig. 1.1). Upon further study, more complex differentiation patterns emerge. For example, it has been observed that cells cultured for a period of time on stiff substrates, such as standard tissue culture polystyrene (TCPS) plates, differentiate into osteogenic lineage cells even after being transferred from the stiff to a softer substrate [167]. Seeding MSCs on a phototunable substrate demonstrates that osteogenic patterns of gene expression persist even after decreasing the stiffness of the substrate [167]. This ”mechanical memory”: the ability of MSCs to remember previous physical stimuli depends on both culture time and substrate stiffness (depicted in Fig. 1.1).

Due to mechanical memory, MSC differentiation in vitro can yield unpredictable (and undesirable) results. Mechanical memory also makes it very difficult to perform certain in vitro assays reliably, for example on extremely soft or stiff substrates, or assays with very long or short incubation periods. Such extreme culture conditions are nonetheless important to assess in order to fully elucidate the relationship between MSC fate and substrate stiffness [125]. In addition to the impracticality of performing short (i.e. seconds) or long (i.e. months) incubation experiments, experimental knock-downs of key genes involved in mechanotransduction, such as Yesassociated protein (YAP), can be lethal or highly toxic in vitro and in vivo [123, 8]. There is thus a need for in silico studies to simulate culture conditions and to map the MSC fate predictions to experimental results describing mechanically induced cell differentiation.

Several mathematical models of mechanotransduction have been built to describe cell differentiation directed by external mechanical stimuli [15, 101]. These include, for example, analysis of the role of *YAP/TAZ*, the transcriptional factors *YAP* and transcriptional co-activator with PDZ-binding motif (*TAZ*), in mechanosensing [140], and models that aim to predict cell differentiation during bone healing [15, 66, 138]. Mousavi et al developed a 3D mechanosensing computational model to illustrate that matrix stiffness can regulate MSC fates. Their simulation results of MSC differentiation in response to substrate stiffness are in agreement with published experimental observations [101]. Burke et al built a computational model to test whether substrate stiffness and oxygen tension regulate stem cell differentiation during fracture healing [15]. Their model predicted the presence of major processes involved with fracture healing, including cartilaginous bridging, endosteal and periosteal bony bridging, and bone remodeling, using parameters related to cell proliferation, oxygen tension, and substrate stiffness. However, these models are limited in that the effects of regulatory factors were not considered [15, 101, 140, 66, 138]. Furthermore, these studies used different models to represent different experimental observations. Hence it is difficult to describe the overall cell state space and to study the transitions between cell fates [15, 101, 140, 66, 138]. Thus, there is a need for a dynamic mathematical model, which can stimulate a continuous range of stiffness values and their associated cell fates.

Here we present a mathematical model of MSC differentiation controlled by the following set of core mechanisms (Fig. 1.2 and Table 1.1) [35, 49, 125]. The MSCs sense the stiffness of their environment directly via their adhesion to the substrate. The transcriptional factors *YAP* and *TAZ* mediate the signal via their interaction with downstream genes involved in cell differentiation. *TUBB3*, a gene encoding Tubulin beta-3 chain tightly correlated with a neurogenic cell fate is expressed when MSCs receive stimuli from a soft stiffness environment (<1 kPa) [35]. *PPARG*, peroxisome proliferator-activated receptor gamma, encodes an adipogenic marker and has been shown to be turned on in soft stiffness environments (∼1 kPa) [49]. *MYOD1*, myogenic differentiation 1, a myogenic gene turned on in medium-stiff environments (∼10 kPa), encodes key factors regulating muscle differentiation [35]. *RUNX2*, runt-related transcription factor 2, an osteogenic gene which is upregulated in high stiffness environments (∼40 kPa), is a key transcriptional factor involved in osteoblast differentiation [35] (Fig. 1.1). We use this set of four lineage-specific genes in our model to minimally describe the transcriptional changes observed during MSC differentiation into four distinct cell fates under the influence of mechanical stimuli mediated by *YAP/TAZ* signaling.

Based on the proposed regulatory network structure (Fig. 1.2), we simulate gene expression dynamics under different mechanical dosings. Each in silico experiment describes MSCs cultured in two passages: a first seeding and a second seeding. The substrate stiffness for the first seeding and the duration of the first seeding are particularly important in cell fate determination of MSCs. We also discover an important role for the second seeding stiffness through our simulation studies. Crucially, this two-seeding setup permits mechanical memory to be observed and studied. We assess when cell fates are determined not only by the current substrate stiffness but also by past exposure and find that a memory region exists for each of the four MSC-derived cell lineages studied. Our model demonstrates that stiffness-based MSC differentiation results from non-cooperative regulation of representative genes. Moreover, we show that lowering the second seeding stiffness of MSCs leads to a more diverse palette of MSC fates.

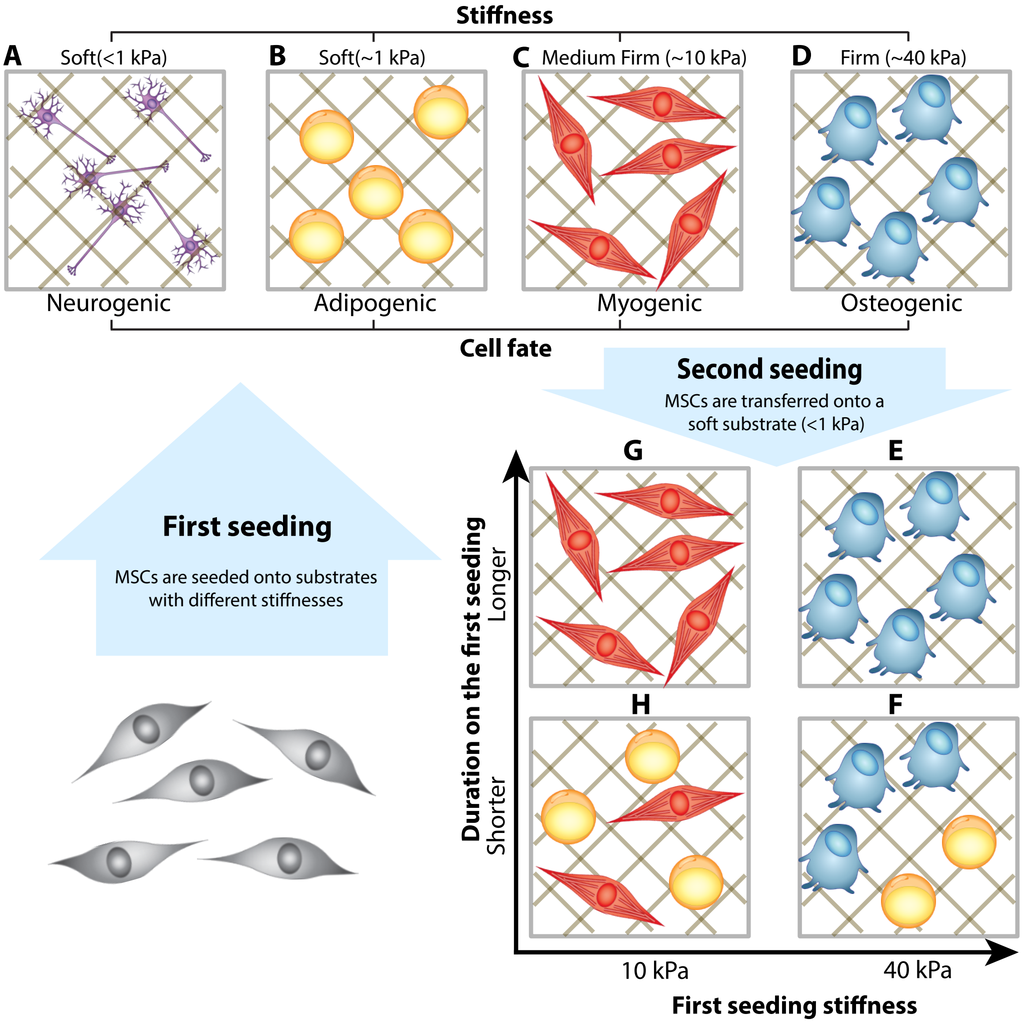


Figure 1.1: Mesenchymal stem cells (MSCs) exhibit mechanical memory. A, B, C, D: MSCs differentiate into distinct lineages under different substrate stiffness conditions by upregulating lineage marker genes *TUBB3* (<1 kPa stiffness, the neurogenic fate), *PPARG* (∼1 kPa stiffness, the adipogenic fate), *MYOD1* (∼10 kPa stiffness, the myogenic fate), or *RUNX2* (∼40 kPa stiffness, the osteogenic fate). When re-seeded onto a soft substrate (∼1 kPa), MSCs are expected to undergo adipogenic differentiation [35, 49, 96]. E, F: However, for higher first seeding stiffness values (>10 kPa), or for long first seeding durations (>10 days), mechanical memory leads to heterogeneous osteogenic differentiation [167]. G, H: The model predicts that for high first seeding stiffness values (∼10 kPa), or for long first seeding durations, mechanical memory leads to heterogeneous myogenic differentiation.

1.2 Results

1.2.1 A mathematical model based on a mechanotransduction network

The following set of biological assumptions has been used to develop the mathematical model. MSCs differentiate according to their surrounding mechanical environment [45, 148, 69, 49, 141]. Directed differentiation towards a particular lineage can be guided if the cells are cultured in a microenvironment that mimics the tissue elasticity of the environment in vivo [45, 148, 141]. Stiff substrates promote cell-ECM adhesion interactions via integrins [49]. These adhesive interactions control the localization of downstream transcriptional factors *YAP* and *TAZ*, which have been identified as mechanical sensors and mediators of such signals [49, 32]. *YAP/TAZ* localizes in the cytoplasm on soft substrates ($\sim$1 kPa) and can relocalize to the nucleus on stiff substrates ($\sim$40 kPa), thus functioning as a mechanosensitive transcription factor [49, 32].

Additionally, *YAP/TAZ* has been reported to be an upstream factor of a number of genes associated with cell differentiation cues [49, 32, 142]. For example, the inhibition of *TUBB3* can be attenuated by *YAP* depletion, whereas that the factor *PPARG* binding to *TAZ* results in inhibition of transcription from the aP2 promoter [56, 57]. *TAZ* functions as an enhancer of MYOD-mediated myogenic differentiation. *RUNX2* can also bind to *TAZ* and cause osteocalcin to be expressed, thus promoting osteogenic differentiation [56, 57]. To describe these interactions, we model *YAP/TAZ* as both a downstream factor of the mechanical stimulus from the ECM and an upstream factor of the selected cell lineage genes [35, 50] (Fig. 1.2 and Table 1.1). Previous references show an intriguing relationship between morphological changes to MSCs and their lineage differentiation potential, whereby morphological changes have been shown to be instrumental to the process of MSC differentiation

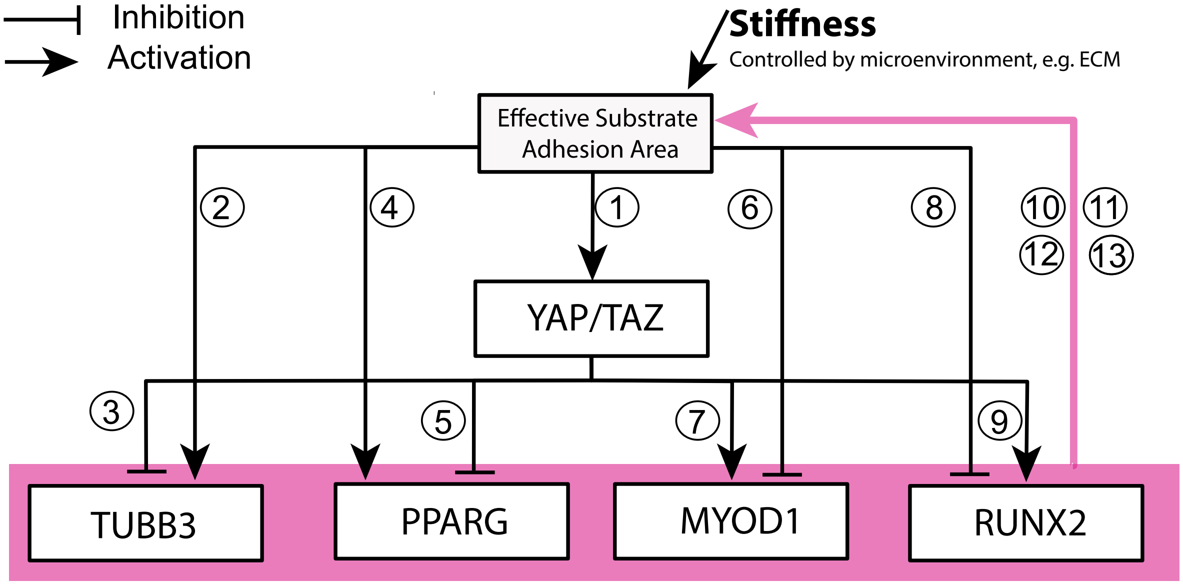

Inhibition
Activation
Stiffness
Controlled by microenvironment, e.g. ECM
Effective Substrate Adhesion Area
YAP/TAZ
1
2
3
4
5
6
7
8
9
10
11
12
13
TUBB3
PPARG
MYOD1
RUNX2

Figure 1.2: Regulatory network used to construct the mathematical model. The boxes represent genes or factors involved in MSC differentiation and the lines with arrows and with bars denote gene activation and inhibition respectively. External stiffness affects the substrate adhesion area. The pink line with an arrow denotes regulations by all species within the pink box. The circled indices refer to experimental evidence for each interaction, details of which are given in Table 1.1

[35, 141, 32, 169, 70, 92]. In particular, it was shown that MSC osteogenic differentiation is enhanced by the morphological change of MSCs and *MYOD1* induced the myogenic differentiation efficiency via the morphological change of MSCs [95, 127]. Other factors regulating cell spreading such as *NKX2.5* were integrated in the model implicitly [30]. Therefore, we model a feedback loop between the lineage-specific target genes and the cellular sensing of substrate stiffness.

In order to predict how mechanical dosing influences MSC differentiation, we use ordinary differential equations to model the MSC lineage regulatory network [1, 116, 134, 26] (Fig. 1.2 and Table 1.1). We assume that changes in the stiffness of the substrate act as stimulus to the

Table 1.1: The references of regulatory interactions in the network

Index of Arrows	**Interactions**	**References**
1	*YAP/TAZ* is identified as mechanical sensors and mediators.	Halder, G et al, 2012; Dupont S. et al. 2011. [49, 32]
3	The inhibition of *TUBB3* can be attenuated by *YAP* depletion.	Alarcon, C et al. 2009 [3]
5	*PPARG* can be bound to *TAZ*, which results in transcription inhibitions from the aP2 promoter.	Hong, J.H. et al, 2006. [57]
7	*TAZ* functions as an enhancer of MYOD-mediated myogenic differentiation.	Jeong, H. et al, 2010. [63]
9	*RUNX2* has binding domain to *TAZ* for osteocalcin expression.	Hong, J.H. et al, 2006. Hong, J.H. et al, 2005 [57, 56]
10,11, 12,13	Increased cell spreading results in higher stiffness sensitivity via increased binding of integrins to the ECM.	Halder G et al, 2012. Sun Y et al, 2012. Bernabe B P et al, 2016. [49, 141, 11]
2,4,6,8	These arrows are necessary for the dynamics of *TUBB3*, *PPARG*, *MYOD1*, and *RUNX2* on all possible stiffness environment since *TUBB3*, *PPARG*, *MYOD1*, and *RUNX2* are expressed only on the super soft stiffness (<1 kPa), the soft stiffness (∼1 kPa), the medium stiffness (∼10 kPa), and the high stiffness (∼40 kPa) environment respectively.	Engler, A.J. et al,2006; Halder G et al, 2012 [35, 49]

Table 1.2: Parameter values of the mathematical model

Index	**Parameter**	**Value**	**Estimated from references**	**Index**	**Parameter**	**Value**	**Estimated from references**
1	k_1	0.2	[35, 49]	2	k_2	2.2	[35, 49]
3	k_3	5	[35, 49]	4	k_4	9	[35, 49, 167]
5	k_5	4	[35, 49]	6	k_6	2.9	[35, 49]
7	k_7	3	[35, 49]	8	k_8	5	[35, 49, 167]
9	k_9	2	[35, 49, 167]	10	K_1	600	[35, 49]
11	n_1	4	[35, 49]	12	K_2	1.1	[35, 49]
13	n_2	2	[35, 49]	14	K_3	1300	[35, 49]
15	n_3	6	[35, 49]	16	K_4	0.8	[35, 49, 167]
17	n_4	2	[35, 49]	18	K_5	20,000	[35, 49]
19	n_5	4	[35, 49]	20	K_6	1	[35, 49]
21	n_6	20	[35, 49]	22	K_7	60,000	[35, 49]
23	n_7	6	[35, 49]	24	K_8	1.1	[35, 49]
25	n_8	20	[35, 49]	26	K_9	0.1	[35, 49]
27	n_9	2	[35, 49]	28	K_{10}	0.5	[35, 49]
29	n_{10}	8	[35, 49]	30	K_{11}	0.89	[35, 49]
31	n_{11}	2	[35, 49]	32	K_{12}	4	[35, 49]
33	n_{12}	8	[35, 49]	34	K_{13}	12	[35, 49]
35	n_{13}	20	[35, 49]	36	K_{14}	3	[35, 49]
37	n_{14}	60	[35, 49]	38	K_{15}	16	[35, 49]
39	n_{15}	45	[35, 49, 167]	40	K_{16}	4.5	[35, 49, 167]
41	n_{16}	55	[35, 49, 167]	$d_i(i = 1, 2, ..., 6)$		1	[35, 49, 167]

cell (mediated by stiffness receptors) [15, 90]. We use Hill functions to model the chemical activation/inhibition [134, 26, 42]. We model the feedback loop that controls mechanical memory via a non-cooperative regulation, i.e., any of the lineage-specific genes (*TUBB3*, *PPARG*, *MYOD1*, *RUNX2*) can increase the effective stiffness adhesion area (we use "OR-GATE" logic). The feedback loop controls the expression of *YAP/TAZ* and its downstream genes via the stimulus (i.e., the change in stiffness [167]). We also test a feedback model of cooperative regulations (where *TUBB3*, *PPARG*, *MYOD1* and *RUNX2* must act together to increase the effective stiffness adhesion area, i.e. "AND-GATE" logic) but find that it does not satisfy the dynamical requirements of the MSC differentiation system (see Methods for full details).

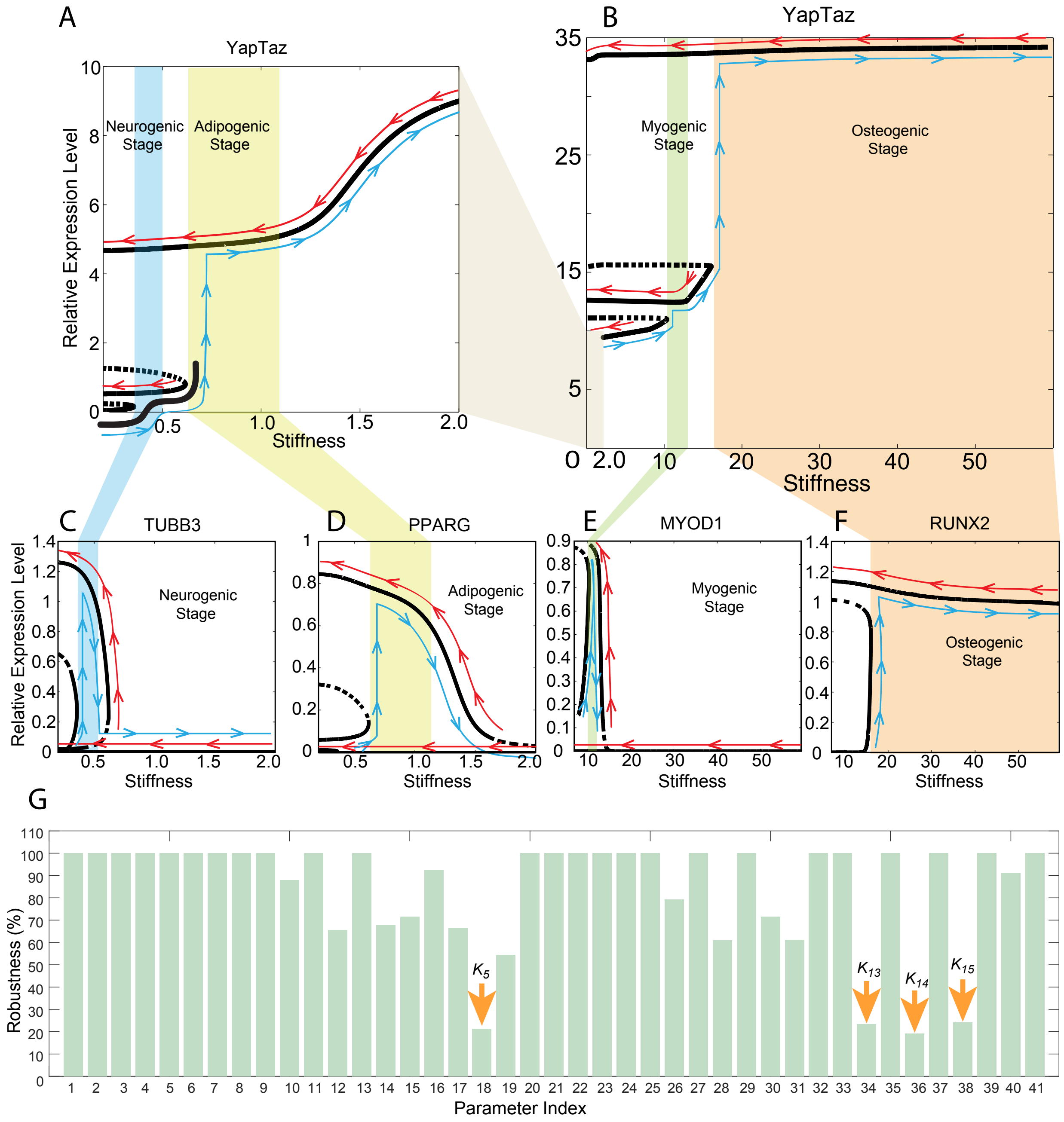


Figure 1.3: Multistability in the MSC differentiation network. The relative expression level of *YAP/TAZ* in a stiffness range from 0.1 kPa to 60 kPa is shown (B), with inset (A). The relative expression levels of lineage-specific genes are shown in (C-F). On each plot the x-axis is the stiffness of the substrate and the y-axis is the relative gene expression level. Blue lines illustrate changes in the relative expression level as the stiffness increases; red lines illustrate changes in the relative expression level as the stiffness decreases. (G). The robustness of the parameters in the mathematical model. The x-axis is the parameter index, corresponding to the notation of Table 1.2. The y-axis is the robustness of the parameters (defined in Methods)

1.2.2 Model simulations predict mechanical memory regions for each lineage-specific gene

The non-cooperative regulation model displays multiple steady states over the behavioral regions that we have investigated (with first seeding stiffness values ranging from 0.1 kPa to greater than 100 kPa; Fig. 1.3). This range is sufficient to encompass all known in vitro studies [35, 49, 167]. In Fig. 1.3A and B the multiple steady states of *YAP/TAZ* expression over the stiffness range studied are shown, and changes in the *YAP/TAZ* state can be visualized as the stiffness increases (blue lines) or decreases (red lines). The nonlinear relationship between *YAP/TAZ* and the stiffness of the substrate along the blue lines is consistent with previous observations [125, 142].

Figure 1.3C demonstrates bistability in the relative gene expression of *TUBB3* (driver of neurogenic differentiation) downstream of *YAP/TAZ. TUBB3* is "OFF" when the stiffness is lower than 0.2 kPa. It will be turned "ON" as the stiffness increases to 0.25 kPa. It turns "OFF" again as the stiffness increases further. Meanwhile, *TUBB3* stays "ON" when the stiffness decreases below 0.2 kPa, thus highlighting the mechanical memory observed during neurogenic differentiation. Notably, *TUBB3* stays "OFF" as the stiffness decreases from 0.6 kPa. We define the region of stiffness from 0.25 to 0.55 kPa as a "differentiation memory region" for *TUBB3.* This means that if the first seeding stiffness is within this range, the cell will "remember" the stiffness of this first seeding substrate, and will differentiate according (towards a neurogenic fate) upon reseeding. Our model also predicts novel differentiation memory regions for *PPARG* (0.6 to 3 kPa; Fig. 1.3D) and *MYOD1* (10 to 15 kPa; Fig. 1.3E). *RUNX2* displays the largest differential memory region of the four lineagespecific marker genes studied. Figure 1.3C-F collectively demonstrate a bistable region for each of the four lineage-specific genes studied. This is a startling prediction: that a region of mechanical memory exists for each of the cell fates, not just for osteogenic differentiation, as has been previously reported [167]. For neurogenic and adipogenic differentiation, the memory regions are smaller than that of osteoblasts yet may still be of great importance for stem cell fate regulation. The true contribution of each will require further study to elucidate, as a host of interacting factors contribute to the neurogenic and adipogenic cell fate decisions, including those which are not currently included in our model, such as the role of substrate-induced stemness and of epithelial to mesenchymal transition [43, 89, 21]. To test the robustness of the mathematical model we calculate the values of the robustness of each parameter in Eqs. (1.1, 1.2, 1.3, 1.4, 1.5, and 1.6) with respect to the memory and multistability of the system (full details of our methodology are in Methods). Out of the 41 parameters tested, 37 are robust to small changes for the majority of perturbations tested (and many of these 37 were robust more than 80% of the time) (Fig. 1.3G). Four parameters are found to be sensitive to small perturbations. All of these four parameters are involved in myogenic or osteogenic differentiation. Both these processes involve relatively large memory regions, thus it is possible that following these perturbations memory is maintained over parts of - but not the entire - original memory regions. Overall, we find that the system displays robustness using the parameters given in Table 1.2, with regard to the memory effects and the multistability of the states.

1.2.3 A lower second seeding stiffness permits a greater number of MSC lineages

Potential energy landscape analysis is an appealing method with which we can investigate the system and study the MSC differentiation propensities under different conditions [156, 9, 155]. Since it is not possible to write down a complete expression for the potential energy of the system, we use an approximate method derived from mean field theory in order to calculate quasi-potential in terms of the six system variables [155, 154]. Explicitly, we calculate the potential of the system as $U(X) = -ln(P_{ss}(X))$, where $P_{ss}(X)$ is the total probability of the state vector X, and X describes all the states of the system [155, 154].

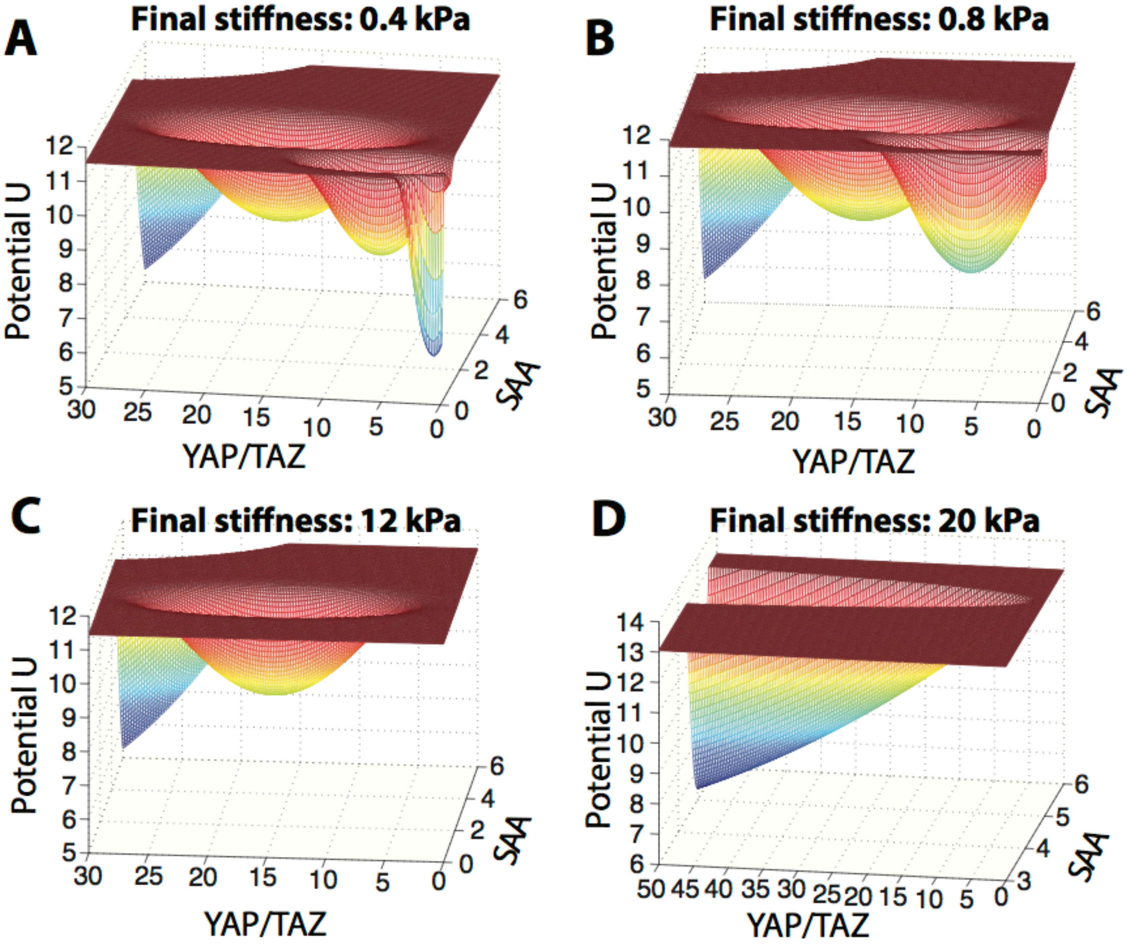


Figure 1.4: Potential landscapes of the regulatory network under different stiffness conditions. In each figure the relative stiffness level (input to the system) is plotted on the x-axis, the relative expression level of *YAP/TAZ* is plotted on the y-axis, the energy potential function U is plotted on the z-axis. Potential energy landscapes are shown with stiffness values of ~0.4 kPa (A), ~0.8 kPa (B), ~12 kPa (C) and ~20 kPa (D)

In order to visualize this potential function we project it onto a two-dimensional plane, defined by the species in our model: *YAP/TAZ*, and the effective stiffness adhesion area (SAA). In doing so we integrate out the four remaining system variables (*TUBB3*, *PPARG*, *MYOD1*, and *RUNX2*) [155, 154]. We are thus able to study how the potential depends on these variables for different stiffness values. In Fig. 1.4 we show the potential functions for four different conditions (we change the second-seeding stiffness values). Overall, we find that by reducing the second seeding stiffness, a greater number of steady states is permitted. We simulate more than 10,000 initial conditions in order to avoid becoming trapped in local minima [155, 154]. We observe that across the entire landscape there are four stable states (or basins of attractions), representing neurogenic, adipogenic, myogenic, and osteogenic cell lineages. At a final stiffness of ∼0.4 kPa, MSCs can differentiate into each of the four possible lineages (Fig. 1.1A). Only at such sufficiently small values for the second stiffness can MSCs differentiate into neurons: the basin of attraction for the neurogenic fate (i.e. the probability of differentiating into a neuron) is the smallest of the four fates. This means that mechanical memory is observed only over a small range of space. In comparison, a much greater mechanical memory effect is seen for the osteogenic lineage, corresponding to a larger basin of attraction. Figure 1.4B and C show the potential landscapes at second seeding stiffness values of ∼0.8 kPa and ∼12 kPa, respectively. The number of basins decreases to three, and then two, as the second seeding stiffness increases. When the second seeding stiffness increases further to ∼20 kPa, we have only one remaining basin of attraction, thus only one possible cell fate: in this region the largest mechanical memory effect is seen, and osteogenic differentiation dominates. These data intriguingly suggest that simply by controlling the substrate stiffness upon re-seeding we can control the number of cell fates that are accessible to MSCs.

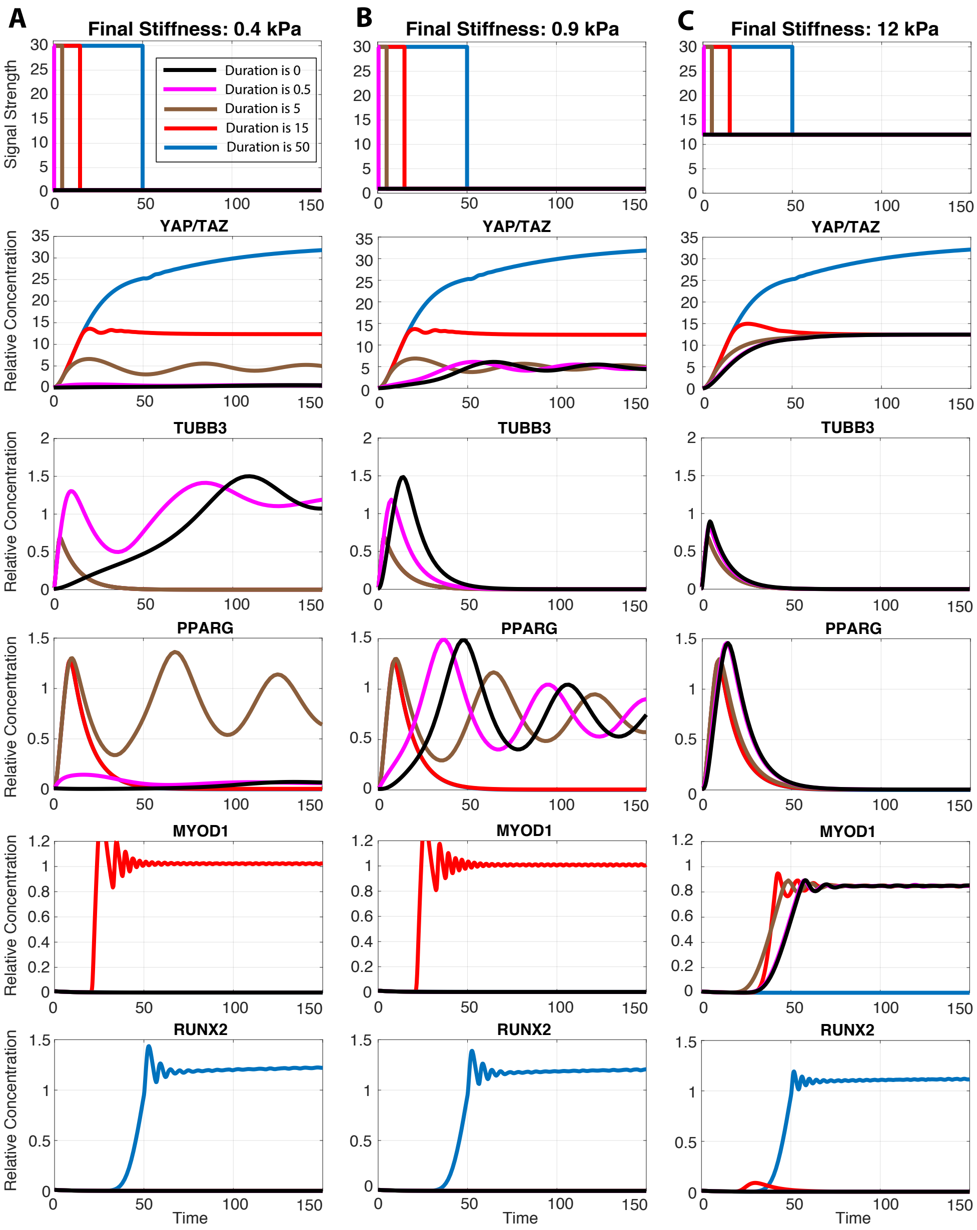


Figure 1.5: The duration of the first seeding regulates MSC fates via mechanical memory. The first seeding stiffness in this figure is 30 kPa. The second seeding stiffness is 0.4 kPa (A), 0.9 kPa (B) or 12 kPa (C). When the duration of the first seeding is 50 (blue lines), MSCs undergo osteogenic differentiation according to memory. When the duration of the first seeding is 15 (red lines), MSCs undergo myogenic differentiation. When the duration of the first seeding is 5 (brown lines in columns A and B), MSCs differentiate into adipocytes or myogenic cells. When the duration of the first seeding is 0.5 (pink lines in column A), MSCs are able to undergo adipogenic, myogenic, or neurogenic differentiation. Finally, when the duration of the first seeding is 0 (black lines), MSCs are able to undergo adipogenic, myogenic, or neurogenic differentiation.

1.2.4 The duration of the initial seeding determines the fate of an MSC

In addition to studying the effect of the second seeding stiffness on the fate of MSCs, we perform tests to assess the agreement between our model and in vitro observations regarding MSC differentiation [35, 32]. Specifically, we manipulate the stiffness of the second seeding substrate and the duration of the first seeding, and find, consistent with previous studies [62, 102], that both of these variables play an important role in the fate determination of an MSC upon differentiation. In addition these simulation results highlight several new phenomena.

In order to examine how the first seeding duration affects MSC fates, we use a non-dimensionalized version of the model, that is, we express time in relative units. In Fig. 1.5A, the first and second seeding stiffness values are 30 kPa and 0.4 kPa, respectively. When the duration of the first seeding time is 50 (blue line), MSCs differentiate into osteoblasts (consistent with [62]): *RUNX2* is the only gene that is highly expressed under this condition. When the first seeding duration is 15 (red line), MSCs differentiate into skeletal muscle cells (*MYOD1* high); when the first seeding duration is five (brown line), MSCs differentiate into adipocytes (*PPARG* high). Finally when the first seeding duration is 0.5 or 0 (pink and black lines), MSCs differentiate into neurogenic cells (*TUBB3* high). These results are consistent with previous studies and highlight the breadth of control that mechanical memory enables: MSCs can be directed to four different fates by changing only the duration of the first seeding, keeping both of the first and the second substrate stiffness values constant. Although mechanical memory is not observed when the first seeding duration is less than 0.5, for the first seeding durations greater than five, we predict that mechanical memory will influence MSC fates, directing MSCs towards myogenic or adipogenic lineages.

Mechanical memory persists when the second seeding stiffness increases, but the number of fates accessible to an MSC decreases, as described in previous sections. In Fig. 1.5B the second seeding stiffness is 0.9 kPa. When the relative duration of the first seeding is 50 (blue line), MSCs differentiate into osteoblasts according to mechanical memory. When the relative duration of the first seeding is 15 (red line), MSCs differentiate into myocytes (again, influenced by memory). When the relative duration of the first seeding is 5, 0.5 or 0, however (brown, pink or black lines), MSCs differentiate into adipocytes: mechanical memory is not present when the second seeding duration is less than 15.

Figure 1.5C shows the dynamics of the system when the second seeding stiffness is 12 kPa. For the longest first seeding duration (blue line), MSCs differentiate into osteoblasts, as above, but when the duration is 15 or lower (red, brown, pink or black lines), MSCs differentiate into myocytes. These data illustrate that as the second seeding stiffness increases, the range of first seeding durations over which mechanical memory is observed decreases, which is consistent with the observation from Yang et al [167]. At a second seeding stiffness of 12 kPa, the memory effect is observed only for osteogenic differentiation, and not for any other lineages. Intriguingly, higher first seeding stiffness values for shorter periods of time might accelerate an MSC towards lineage commitment. *TUBB3* expression approaches the steady state quickly following stimulation on a 30 kPa substrate for a relative time of 0.5 (Fig. 1.5A, pink line). Compare this to the differentiation characteristics of an MSC seeded only on a 0.3 kPa substrate (Fig. 1.5A, black line); the latter takes a longer time to differentiate.

1.2.5 Feedback signaling onto the effective substrate adhesion area

Mechanotransduction pathways may contain positive feedback loops in which integrin engagement activates actomyosin cytoskeleton contractility, resulting in morphological changes affecting the adhesion area of the substrate [35, 141, 32, 169, 70, 92, 95, 127]. Here we assess the importance of such feedback. Figure 1.6 shows the relative expression levels of the

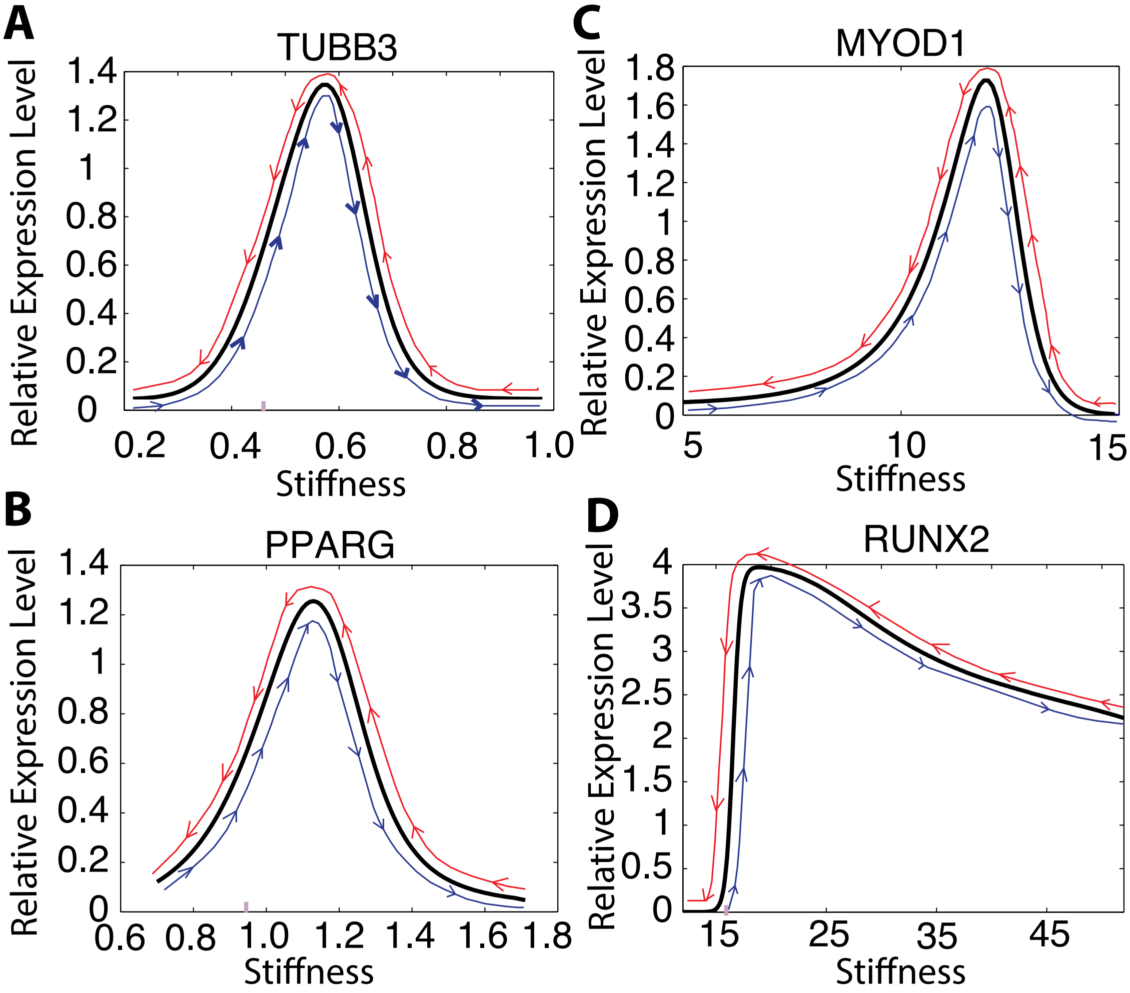


Figure 1.6: The MSC network precludes multistability when feedback loops are blocked. Shown are the steady states of *TUBB3* (A), *PPARG* (B), *MYOD1* (C), and *RUNX2* (D) under different stiffness values. In each figure the x-axis denotes the stiffness and the y-axis denotes the relative expression levels of specific lineage genes at steady states (black lines). The blue lines illustrate how the relative gene expression at the steady state changes as the stiffness increases. The red lines illustrate how the relative gene expression level at the steady state changes as the stiffness decreases

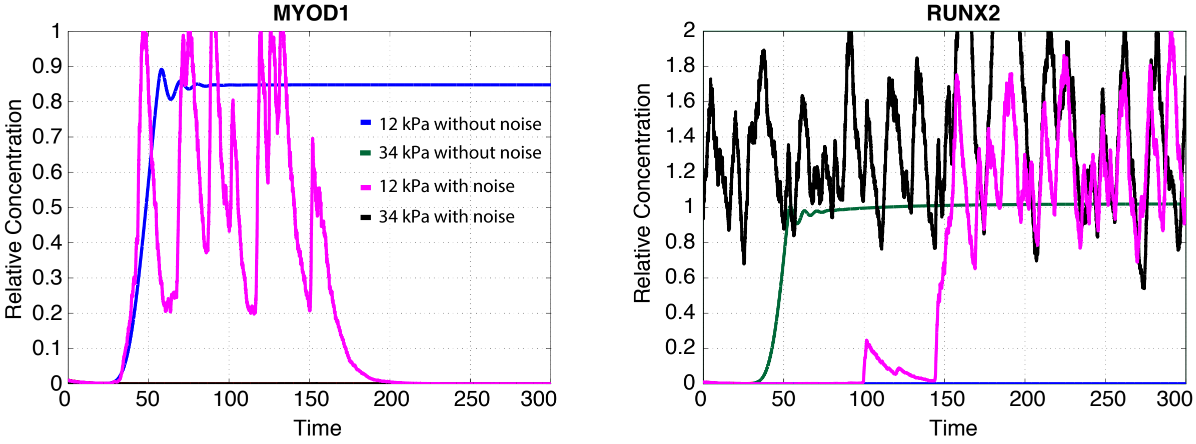


Figure 1.7: Stochastic gene expression dynamics under different stiffness conditions. The green and blue lines depict the relative expression levels of genes from the deterministic model. The magenta and black lines depict the relative expression levels of genes from the stochastic differential equation model with noise term $\sim N(0, 0.05)$. Blue and magenta lines represent a first-seeding stiffness of 12 kPa, green and black lines represent a first-seeding stiffness of 34 kPa. The final seeding stiffness is 12 kPa in all cases

lineage-specific genes at steady states for a range of substrate stiffness values. In Fig. 1.6A, we block the feedback from *TUBB3* onto the effective substrate adhesion area. We see that the bistability that was observed in Fig. 1.3 is no longer present: no hysteresis effect can be seen when the substrate stiffness is increased or decreased (illustrated by the blue and red lines). Thus, no mechanical memory effect remains for *TUBB3* during MSCs differentiation. Similar results are obtained for *PPARG* (Fig. 1.6B), *MYOD1* (Fig. 1.6C) and *RUNX2* (Fig. 1.6D) when the final seeding stiffness is 0.9 kPa, 10 kPa and 16 kPa, respectively. The mechanical memory of the genes disappears when the feedback loops are removed. Collectively our simulation results illustrate that the feedback loops downstream of the stiffness of substrates are necessary for the mechanical memory.

1.2.6 Noise can induce fate switching during MSC differentiation

There is inherent noise in gene expression dynamics [34, 17]. We employ a stochastic differential equation (SDE) model (described in Methods) to study the effects of gene expression noise on MSC differentiation [20, 55]. We find that SDE simulations broadly recapitulate the results obtained in the deterministic case, however under certain conditions fate switching is observed. In Fig. 1.7 we simulate a system of SDEs based on the deterministic model with multiplicative noise added to the expression level of each gene; blue and dark green lines describe the relative gene expression under the deterministic model, while pink and black lines describe analogous results under the SDE model. We vary the initial seeding stiffness while keeping the second seeding stiffness constant at 12 kPa. In the deterministic case, we see that *MYOD1* is expressed when the value of the initial stiffness is 12 kPa, and not when the value is 34 kPa. Conversely, *RUNX2* is not expressed at an initial stiffness of 12 kPa, but is expressed when the initial stiffness is 34 kPa: here stem cells are differentiating according to mechanical memory. In the stochastic case, a different picture emerges. First we note that the memory effect observed for osteogenic differentiation in the deterministic case (driven by *RUNX2* expression) is preserved under the stochastic model (Fig. 1.7 black line). However, in the stochastic case, at 12 kPa, *MYOD1* is expressed transiently: as its expression declines to zero, *RUNX2* is turned on. Thus noise has induced a fate transition between myogenic and osteogenic lineages. At 34 kPa no such transitions are observed: *RUNX2* is expressed constitutively.

1.3 Discussion

Mesenchymal stem cell fate can be controlled by mechanical dosing [35]. Mechanical memory (past mechanical dosing) also affects stem cell fate, particularly when the initial substrate is stiff [167], it is difficult however to experimentally test the effects of mechanical memory over a wide range of culture conditions. Here we have presented a mathematical model that allows such tests to be performed, producing several striking predictions. We first assessed whether the model is able to recapitulate experimental studies, and find that it does agree with evidence showing MSC differentiation into neurons or adipocytes on softer substrates, and myocytes or osteoblasts on stiffer substrates. We then analyzed model behavior over longer timescales, and found that a mechanical memory region exists for each of these MSC-derived cell lineages, with substantial variation in the memory stiffness range for each cell fate. Previously, a memory region has only been observed during osteogenic differentiation, and even then, only qualitative assessment of its behavior was made. We are able to provide bounds on the substrate stiffness ranges permissive of memory effects for all four lineages.

Upon re-seeding MSCs onto a second substrate, the stem cells differentiate according to mechanical memory under certain conditions. We predict that (in addition to the stiffness of the first substrate) the duration of the first seeding also directly influences stem cell memory. By changing only the duration of the initial seeding we can directly influence cell fate. The number of fates accessible to the MSC can also be controlled by the final seeding stiffness. Landscape analysis demonstrates that, for a constant first seeding stiffness and duration, a higher second seeding stiffness limits the number of MSC fates accessible, and that a sufficiently low final seeding stiffness is permissive of differentiation into all four cell fates. We also found that the feedback loop connecting lineage-specific genes to the effective surface adhesion area is critical for the mechanical memory of MSC differentiation. This might be due to integrin-substrate binding, or morphological changes that occur upon differentiation [35, 148, 160, 141].

As well as their direct relevance for in vitro studies, our model predictions also have important implications for the design of regenerative therapeutics. A major challenge here is lack of precision in cell fate control following transplantation. A better understanding of the relationship between mechanical conditions, culture duration, and stem cell fates is needed. By defining the substrate stiffness limits that regulate MSC fates, this study provides means to design experimental protocols that constrain cells to be confined within fate boundaries, thus avoiding differentiation towards an undesirable fate [23, 97, 112, 79]. Mechanical memory could be employed advantageously here, e.g. by preconditioning MSCs via mechanical dosing. An improved understanding of the MSC mechanotransduction pathway will also affect our ability to control multipotency, and should enable us to better culture undifferentiated MSCs in vitro.

In order to study additional effects of the mechanotransduction pathway on stem cell fate, a model that describes a larger regulatory network is needed. Cell-cell interactions have not yet been incorporated into our model, although there is a large body of work detailing the importance of the microenvironment (i.e. the effects of cell-cell interactions and of the niche) on stem cell differentiation [116, 41]. In addition, we have chosen a small set of four lineagespecific genes in order to minimize the size of the model parameter space. Clearly a greater number of genes are involved in the regulation of MSC fate; without a description of this larger transcriptional network we will not be able to describe nuances of mechanically-induced MSC fate dynamics. However, we believe that the dynamics and the attractors corresponding to differentiated cell states observed here constitute core pathway mechanisms that would still underlie cell fate decisions in a larger network.

1.4 Conclusions

In this study we sought to investigate the mechanisms of control exerted via mechanical forces upon mesenchymal stem cells during culture and differentiation. Simulations of the gene expression dynamics under different mechanical dosing conditions have led to several predictions. We found that non-cooperative gene regulation is the most plausible mechanism to describe MSC differentiation and we predict that mechanical memory is a general mechanism affecting all of the MSC-derived lineages in this model. We found that the duration of the initial culture and the substrate stiffness during this initial culture are particularly crucial in determining the MSC fates. In addition, we were able to show that a lower final-seeding substrate stiffness permitted a greater number of MSC fates.

Through careful analysis, the ever-expanding body of high-throughput transcriptomic data will enable the study of ever-more complex gene networks. Both the MSC fate transcriptional network structure and the dynamics of the network need to be inferred from data. Spatial interactions, e.g. arising from niche-mediated effects on MSCs, may necessitate a move towards a suitable model framework such as partial differential equations or cell-based (e.g. Cellular Potts) models. Once a clearer picture emerges, it will be possible to extend our model with the incorporation of relevant new signaling interactions. In doing so, we hope to provide further insight into the complex networks of regulation underpinning mesenchymal stem cell fate.

1.5 Methods

1.5.1 A dynamical model of mesenchymal stem cell fate

We model a simplified gene regulatory network that underpins MSC fate with ordinary differential equations (ODEs) [134, 26].

$$
\begin{aligned}
\frac{d[SAA]}{dt} = {} & \underbrace{k_1 \frac{(S/K_1)^{n_1} + ([TUBB3]/K_2)^{n_2}}{1 + (S/K_1)^{n_1} + ([TUBB3]/K_2)^{n_2}}}_{10} + \underbrace{k_2 \frac{(S/K_3)^{n_3} + ([PPARG]/K_4)^{n_4}}{1 + (S/K_3)^{n_3} + ([PPARG]/K_4)^{n_4}}}_{11} + \\
& \underbrace{k_3 \frac{(S/K_5)^{n_5} + ([MYOD1]/K_6)^{n_6}}{1 + (S/K_5)^{n_5} + ([MYOD1]/K_6)^{n_6}}}_{12} + \underbrace{k_4 \frac{(S/K_7)^{n_7} + ([RUNX2]/K_8)^{n_8}}{1 + (S/K_7)^{n_7} + ([RUNX2]/K_8)^{n_8}}}_{13} - d_1[SAA],
\end{aligned}
\tag{1.1}
$$

$$
\frac{d[YAPTAZ]}{dt} = \underbrace{k_5[SAA]}_{1} - d_2[YAPTAZ] \tag{1.2}
$$

$$
\frac{d[TUBB3]}{dt} = \underbrace{k_6 \frac{([SAA]/K_9)^{n_9}}{1 + ([SAA]/K_9)^{n_9} + ([YAPTAZ]/K_{10})^{n_{10}}}}_{2,3} - d_3[TUBB3], \tag{1.3}
$$

$$
\frac{d[PPARG]}{dt} = \underbrace{k_7 \frac{([SAA]/K_{11})^{n_{11}}}{1 + ([SAA]/K_{11})^{n_{11}} + ([YAPTAZ]/K_{12})^{n_{12}}}}_{4,5} - d_4[PPARG], \tag{1.4}
$$

$$
\frac{d[MYOD1]}{dt} = \underbrace{k_8 \frac{([YAPTAZ]/K_{13})^{n_{13}}}{1 + ([SAA]/K_{14})^{n_{14}} + ([YAPTAZ]/K_{13})^{n_{13}}}}_{6,7} - d_5[MYOD1], \tag{1.5}
$$

$$
\frac{d[RUNX2]}{dt} = \underbrace{k_9 \frac{([YAPTAZ]/K_{15})^{n_{15}}}{1 + ([SAA]/K_{16})^{n_{16}} + ([YAPTAZ]/K_{15})^{n_{15}}}}_{8,9} - d_6[RUNX2], \tag{1.6}
$$

Where S and $[SAA]$, are the relative levels of the stiffness (input to the system) and of the effective stiffness adhesion area, respectively. $[YAPTAZ]$, $[TUBB3]$, $[PPARG]$, $[MYOD1]$, and $[RUNX2]$ denote the relative concentrations of *YAP/TAZ*, *TUBB3*, *PPARG*, *MYOD1*, and *RUNX2*. Since concentration and time in the model are given in relative units, i.e. are dimensionless, then all parameters in the above equations are also dimensionless. d_i ($i = 1, 2, ..., 6$) in Eqs. (1.1, 1.2, 1.3, 1.4, 1.5, and 1.6) are the degradation rates of the corresponding genes/factors. The terms denoted by the label (1, 2, ..., 9) under the brackets in Eqs. (1.1, 1.2, 1.3, 1.4, 1.5, and 1.6) are the active/inhibitive regulations acting on $[SAA]$, $[YAPTAZ]$, $[TUBB3]$, $[PPARG]$, $[MYOD1]$, and $[RUNX2]$, where the numbers in rectangle boxes are consistent with the circled indices shown in Fig. 1.2 [120]. All values of parameters in Eqs. (1.1, 1.2, 1.3, 1.4, 1.5, and 1.6) shown in Table 1.2 are estimated or approximated according to the behaviours that we sought to describe. Parameters values are fit to qualitative features of the biological system [35, 49, 167, 125, 142] (See Appendices: A Additional file for Chapter 1). The data required performing full inference of the parameters are as-yet unavailable, however the results of our sensitivity analysis show that the models results do not depend crucially on specific values of parameters of the model.

1.5.2 Cooperative regulation model

The terms (10, 11, 12, 13) in Eq. (1.1) are based on the noncooperative regulations of MSCs stiffness sensing. Meanwhile, we model the regulations as the cooperative one and Eq. (1.1) is rewritten below [61].

$$\frac{[SAA]}{dt} = \underbrace{k_1 \frac{(S/K_1)^{n_1} + ([TUBB3]/K_2)^{n_2} + ([PPARG]/K_3)^{n_3} + ([MYOD1]/K_4)^{n_4} + ([RUNX2]/K_5)^{n_5}}{1 + (S/K_1)^{n_1} + ([TUBB3]/K_2)^{n_2} + ([PPARG]/K_3)^{n_3} + ([MYOD1]/K_4)^{n_4} + ([RUNX2]/K_5)^{n_5}}}_{10,11,12,13} - d_1[SAA] \quad (1.7)$$

Rehfeldt et al showed the switch-like nonlinear relationship between S and SAA expanding from 0.5 kPa to much large stiffness (>60 kPa) and *TUBB3*, *PPARG*, *MYOD1*, and *RUNX2* are turned on in their specific ranges of stiffness, which are relatively disjoint [120, 61, 125]. In particular, the stiffness range for the myogenic differentiation is far away from the one for adipogenic differentiation. Based the properties of the system, we can rewrite our model into four different submodels under the corresponding stiffness ranges. They are shown as follows.

$$\frac{[SAA]}{dt} = k_1 \frac{([S]/K_1)^{n_1} + ([TUBB3]/K_2)^{n_2}}{1 + ([S]/K_1)^{n_1} + ([TUBB3]/K_2)^{n_2}} - d_1[SAA], \tag{1.8}$$
$$\frac{[SAA]}{dt} = k_1 \frac{([S]/K_1)^{n_1} + ([PPARG]/K_3)^{n_3}}{1 + ([S]/K_1)^{n_1} + ([PPARG]/K_3)^{n_3}} - d_1[SAA], \tag{1.9}$$
$$\frac{[SAA]}{dt} = k_1 \frac{([S]/K_1)^{n_1} + ([MYOD1]/K_4)^{n_4}}{1 + ([S]/K_1)^{n_1} + ([MYOD1]/K_4)^{n_4}} - d_1[SAA], \tag{1.10}$$
$$\frac{[SAA]}{dt} = k_1 \frac{([S]/K_1)^{n_1} + ([RUNX2]/K_5)^{n_5}}{1 + ([S]/K_1)^{n_1} + ([RUNX2]/K_5)^{n_5}} - d_1[SAA] \tag{1.11}$$

The difficulty is to determine the values of K1. If K1 is less than 1000, the hill function in Equation (1.7) is saturated for high stiffness levels (> 10,000) and it means that the models cannot distinguish the myogenic differentiation and osteogenic differentiation since Eqs. (1.10 and 1.11) both approach the limit $\frac{d[SAA]}{dt} = k_1 - d_1[SAA]$. If K_1 is greater than 10,000, then the model cannot describe the system for low stiffness levels (< 1000) with that *TUBB3* and *PPARG* cannot express under the low stiffness levels since Eqs. (1.8 and 1.9) will respectively approach the limit:

$$\frac{[SAA]}{dt} = k_1 \frac{([TUBB3]/K_2)^{n_2}}{1 + ([TUBB3]/K_2)^{n_2}} - d_1[SAA], \tag{1.12}$$
$$\frac{[SAA]}{dt} = k_1 \frac{([PPARG]/K_3)^{n_3}}{1 + ([PPARG]/K_3)^{n_3}} - d_1[SAA], \tag{1.13}$$

Thus the cooperative regulation model is unable to accurately describe the MSC differentiation system over the range of stiffness values considered.

1.5.3 Sensitivity analysis

In order to calculate the sensitivities of the parameters shown in Table 1.2 with respect to the memory and multistability of the system, we sample 1000 values between 0.2 kPa and 42 kPa; they are taken as the stiffness of the system and they are vectorized as the stiffness vector S_b. We then calculate the steady states, Q_b^{Upper} and Q_b^{Lower}, corresponding to the steady states on the lower bifurcation branch (indicated by blue arrowhead lines in Fig. 1.3C-F, and to the steady states on the upper bifur- cation branch (indicated by red arrowhead lines in Fig. 1.3C-F) for each of the genes: *TUBB3*, *PPARG*, *MYOD1*, and *RUNX2*, using the parameters in Table 1.2. In order to calculate the sensitivity of each parameter, we perturbe it 1000 times under the constraint of CV(coefficient of variance) = 0.05, and calculate the steady states Q_P^{Upper} (with the same initial conditions as Q_b^{Upper}), and Q_P^{Lower} (with the same initial conditions as Q_b^{Lower}). We perform such comparisons for each of the four genes for a total of 41 parameters and 1000 perturbations, thus for the parameter vector $P_i^j(i = 1, 2, ..., 41; j = 1, 2, ..., 1000)$, i.e. the j-th perturbation of the i-th parameter. We count the number (N_i) of P_i^j that satisfies $||Q_P^{Upper} - Q_b^{Upper}||_2 + ||Q_P^{Lower} - Q_b^{Lower}||_2 < TOL$.The tolerance, TOL, is set such the perturbed parameter vector gave rise to the same number of steady states as for the unperturbed case (i.e. multistability and the memory effect is maintained; we set $TOL = 4$). The robustness R_i of the i-th parameter is defined as $\frac{N_i}{10}\%$ and the sensitivity S_i of the i-th parameter is $1 - \frac{N_i}{10}\%$. The robustness values for each of the 41 parameters are shown in the bar graph (See Fig. 1.3G) and the index of the parameters in the graph is consistent with the one in Table 1.2. Four of them are sensitive than the rest and they are marked by yellow arrows in the following bar graph.

1.5.4 Steady state analysis

We compute the steady states of the dynamical system under different S in Eqs. (1.1, 1.2, 1.3, 1.4, 1.5, and 1.6). Here we use the continuation method to compute the steady states and their branches [5, 29].

1.5.5 Landscape potential using a mean field self-consistent approximation and Gaussian approximation

Here we derive an approximation for the potential energy of the system. Starting from the Fokker-Planck equation, we calculate the steady state probability distributions using a self-consistent mean field method [86, 152, 83]. The probability function $P(X,t)$ satisfies the following diffusion equation:

$$\frac{\partial P(X,t)}{\partial t} = -\frac{\partial}{\partial X}[F(X,S)P(X,t)] + D\frac{\partial^2}{\partial X^2}[d(X)P(X,t)] \tag{1.14}$$

where $F(X,S)$ and $d(X)$ are the drift and diffusion part respectively and the noise is weak,i.e. $D << 1$.Note that X is a vector of species ([SAA],[YAPTAZ],[TUBB3],[PPARG],[MYOD1],[RUNX2]) but we have dropped the arrow notation for convenience below. We factor the original probability function using the self-consistent mean field approach [130], $P(X,t) = \prod_{i=1}^{n} P(X_i,t)$ to reduce the computational complexity of solving the original equation on the probability, similar to a previous study [152]. We use the Gaussian distribution to approximate the true distribution [152],leading to a description for the mean and variance of the gene expression:

$$\bar{X}'(t) = F(\bar{X}(t), S) \tag{1.15}$$

$$\sigma'(t) = \sigma(t)A^T(t) + A(t)\sigma(t) + 2D\bar{X}(t) \tag{1.16}$$

where $\bar{X}(t)$ is the mean value of $X(t)$, (t) is the variance matrix, the matrix element $\alpha_{ij}(t)$ of $A(t)$ is $\frac{\partial F_i(\bar{X}(t))}{\partial \bar{X}_j(t)}$, i.e. A is the Jacobian matrix.

Since we consider the steady states, then we need to compute $\bar{X}^{(j)}(\infty)$ and $\sigma^{(j)}(\infty)$ from $\bar{X}'(t) = 0$ and $\sigma(t) = 0$, for $j = 1, 2, ..., m$ respectively, where m is the number of basins of attraction. We consider only diagonal elements of $\sigma^{(j)}(\infty)$ from mean field splitting approximation. For each variable $\bar{X}_i^{(j)}(\infty)$, the probability distribution can be estimated using the mean and variance and based on Gaussian approximation [152, 58].

$$P^{(j)}(X_j, \infty) = \frac{1}{\sqrt{2\pi\sigma^{(j)}(\infty)}} exp[-\frac{[X_i \bar{X}_i^{(j)}(\infty)]^2}{2\sigma^{(j)}(\infty)}] \tag{1.17}$$

If $m = 1$, we can use Eq. (1.17) to compute the probability distribution of the single basin of attraction. If $m > 1$, then the system permits multistability, and for each basin of attraction we compute its probability distribution. The probability function thus becomes a weighted sum of the probabilities given for each basin of attraction,

$$P^{(j)}(X_j, \infty) = \sum_{j=1}^{m} \omega_j P^{(j)}(X_i, \infty) \tag{1.18}$$

where ω_j is the weighting coefficient of the j-th basin. Assume m attractors, then the number of simulations that end up in each attractor is $N_1, N_2, ..., N_m$. The weighting coefficient for the j-th basin is then calculated as $\omega_j = N_j / \sum_{i=1}^{m} N_i$. Finally, we calculate the potential landscapes based on $U(X) = -lnP(X, \infty)$ [84, 82].

1.5.6 A stochastic differential equation model

A stochastic differential equation (SDE) model for the regulatory network can be constructed via the addition of a noise term [34, 20, 55, 4]:

$$dX(t) = F(X(t), S)dt + \eta X(t) dW(t) \tag{1.19}$$

where W(t) denotes the scalar white noise (or Wiener process), and η is the noise coefficient.

Chapter 2

Identify Cellular States and Their Transitions using Single-Cell Data

[Chapeter 2 is reprinted with the permission from Tao Peng, Qing Nie. SOMSC: Self-Organization-Map for High-Dimensional Single-Cell Data of Cellular States and Their Transitions. bioRxiv, 2017. ©2017 The Authors.[115]]

2.1 Introduction

Heterogeneity of cell populations is considered functionally and clinically significant in normal and diseased tissues, and transitions among different subpopulations of cells, such as differentiation, play critical roles during development and disease recurrence [150, 161, 128]. In recent years, single-cell gene expression profiling technologies are emerging as increasingly important tools in dissecting heterogeneity and plasticity of cell populations in addition to analyzing cell-to-cell variability on a genomic scale [129]. For example, mammalian preimplantation development was analyzed from oocyte stage to morula stage in both human

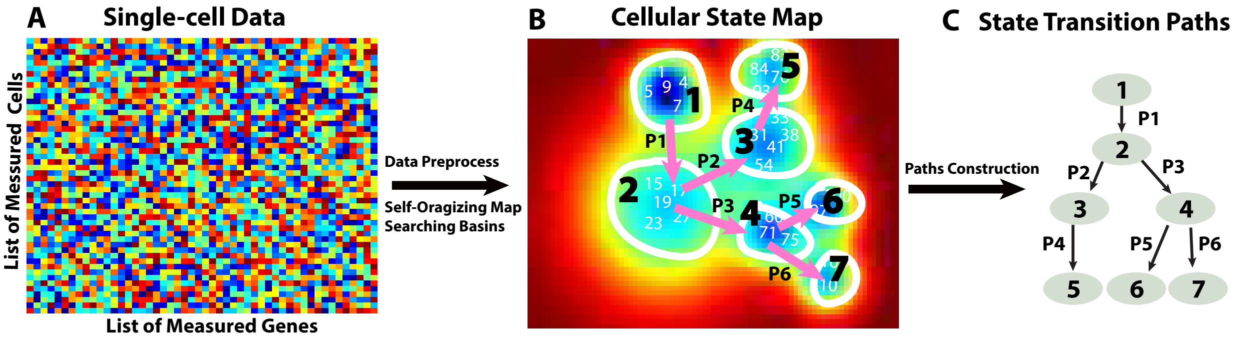


Figure 2.1: A schematic diagram on constructing cellular state maps (CSMs) and transition paths using the SOMSC method. (A) The gene expression data of single cells. (B) A CSM is constructed by the SOMSC using the data. In the CSM each cell is indexed by a number based on a particular given order or a temporal stage at which the data are collected in measurements. A basin of an attraction in the CSM corresponds to one cellular type. The transitions among different cellular states are labeled by arrows such as P1, P2,. . . , and P5. (C) The cellular state lineage trees or differentiation processes are then summarized based on the transition path arrows in the CSM.

and mouse using single-cell RNA sequencing to identify stage-specific transcriptomic dynamics [165, 166]; in breast cancer, gene expression profiles of tumor subpopulations along a spectrum from low metastatic burden to high metastatic burden were obtained using qPCR at the single-cell level [76]; and multiple new phenotypes in healthy and leukemic blood cells were defined using gene expression signatures through analysis of single-cell data [80].

Distinguishing or clustering measured cells computationally through their transcriptomic data (e.g. gene expression) is challenging. The number of cells collected in experiments with successful outputs is usually small whereas the number of genes measured usually is significantly larger [65]. In addition, a group of cells collected at one temporal point from one sample may not be perfectly ordered in time compared to the cells collected at slightly different temporal stages, due to cell-to-cell variability in sampling and its nature of unsynchronized cell divisions [53, 27]. As a result, a pseudo-temporal ordering of single cells in a high-dimensional gene expression space was introduced [147]. The difficulty in analyzing single-cell data becomes particularly evident for systems of differentiation in which new cell types emerge as time advances, such as the cases of lineage progression during development of murine lung [149] and the differentiation trajectory of skeletal muscles [146].

Ordering single cells temporally, grouping cells of similar transcriptomic profiles, finding transition points, and determining branches are among the key steps in analyzing single-cell data. Clustering methods based on Principle Component Analysis (PCA) or Independent Components Analysis (ICA), such as MONOCLE algorithm [146], group cells according to their specific properties of interests. Several other clustering-based methods such as SPADE [122], t-SNE [151], and viSNE [6] were introduced to identify subpopulations within measured cells without an explicit temporal ordering of the cells. In the Wanderlust algorithm [10], a pseudo-temporal ordering technique incorporated the continuity concept in branching processes, however, with an assumption that cells consist of only one branch during differentiation. To address potential nonlinearity of branching processes in differentiation, a diffusion map technique was adapted to single-cell data by adjusting kernel width and inclusion of uncertainties, enabling a pseudo-temporal ordering of single cells in a high-dimensional gene expression space [47]. With a focus on modeling dynamic changes associated with cell differentiation, a bifurcation analysis method (SCUBA) was developed to extract lineage relationships [93].

Meanwhile, a Waddington landscape of gene expression has been widely used to provide a global and physical view in understanding stem cells and cell lineages [44]. In constructing such landscape, a forward stochastic modeling approach is usually applied to a small gene network with an ”energy” function computed through probability density functions or stochastic samplings [38, 171, 172, 18, 84]. In this approach, the prior knowledge of the gene regulatory network needs to be known and the landscape is calculated without dimension reduction in the gene space. However, due to computational cost associated with sampling solutions of stochastic differential equations or solving equations of probability density functions of the gene states, the size of network in the landscape calculation usually is small [156].

Here, we propose a new method to analyze single-cell gene expression data by combining a learning method in an artificial neural network (ANN) and a concept similar to a landscape of gene expression data. In this approach, high dimensions of single-cell data are first reduced to two dimensions through a classical unsupervised learning ANN method: the self-organization map (SOM) [71] in which the topological properties of the input data are preserved through a neighborhood function. A cellular state map (CSM) is then derived to mimic a landscape of gene expression data based on a U-matrix calculated by the SOM. The CSM consists of basins of attractions, which correspond to cellular states, and barriers that separate the different states to indicate directions of transitions between cellular states. Transition paths among the cellular states naturally lead to a pseudo-temporal ordering of the cells. To study effectiveness and capabilities of the method, we apply the self-organization-map for single-cell data (SOMSC) to a set of simulated data and four experimental data sets based on qPCR or RNA-seq collected for systems of cell lineages or differentiation.

2.2 Methods

2.2.1 Preprocess the data

Single-cell gene expression levels measured by qPCR or RNA-seq are prone to having missing values, causing bias in analysis without any preprocessing [13]. In this study, we first remove samples that have many zero values in gene expression data. Specifically, the samples of more than 10% of the total number of genes with missing values will not be used; then the missing values of genes in the rest samples are set to the mean value of that gene at its corresponding stage. Another important step in preprocessing is to normalize the data. Because the SOM algorithm uses the Euclidian distance between gene expression vectors of two samples [121], two genes with drastically different ranges of expression values (e.g. expression values of one gene in $[0, 100]$ whereas the ones of another gene in the range of $[0, 0.1]$) may influence the SOM unfaithfully, as the larger component may dominate the calculation, introducing bias in analysis. Next we normalize the data linearly such that the variance of each gene is equal to one [121]. The normalized data is stored in a matrix in which each row represents expression values of all genes in one single cell, and the number of rows corresponds to the number of single cells in the data after the preprocessing (Figure 2.1A).

2.2.2 Calculate the U-matrix using the Self-Organizing Map

A Self-Organizing Map (SOM) is an effective way of analyzing topology of high-dimensional data, and it projects the data to a low-dimensional surface through a rectangular, a cylinder, or a toroid map [71]. In the SOM, regression of an ordered set of model vectors $m_i \in \mathbf{R}^n$ is made into the space of observation vectors $x \in \mathbf{R}^n$ through the following processes:

$$m_i(t+1) = m_i(t) + h_{c(x),i}(x(t) - m_i(t)), \tag{2.1}$$

where t is an index for a regression step. A regression procedure is performed recursively for each sample $x(t)$. The scalar multiplier $h_{c(x),i}$ is a neighborhood function, acting like a smoothing or blurring kernel over computational grids in the SOM, and often takes a form of Gaussian:

$$h_{c(t),i} = \alpha(t)exp(-\frac{||r_i - r_c||^2}{2\sigma^2(t)}), \tag{2.2}$$

where $0 < \alpha(t) < 1$ is a learning-rate factor, which decreases monotonically through regression steps; $r_i \in \mathbf{R}^2$ and and $r_c \in \mathbf{R}^2$ are locations in the computational grids, and $\sigma(t)$ corresponds to the width of the neighborhood function that also decreases monotonically in each regression step. The subscript $c = c(x)$ is obtained when the following condition is achieved:

$$\forall i, ||x(t) - m_c(t)|| \leq ||x(t) - m_i(t)||. \tag{2.3}$$

Consequently, $m_c(t)$ is the "winner" which matches the best with $x(t)$. The comparison metric $|| \bullet ||$ is selected as the Euclidean metric in Eq.(2.3,2.2). If there are multiple $c(t)$ satisfying Eq. (2.2) with discrete-valued variables, $m_c(t)$ is selected at random for the winner. In the method, a toroid map is used in order to reduce edge effects of the data on the overall mapping [153]. Applying the SOM to the normalized single-cell gene expression data leads to a unified distance matrix (U-matrix) U, representing distances between neighboring map units [71].

2.2.3 Trace the lineage trajectory

Construct Cellular State Map (CSM)

To investigate structure of high-dimensional gene expression data, we first define a cellular state map (CSM) M_{cs} based on the U-matrix U through the equation:

$$M_{cs} = \frac{1}{1 + e^{-\gamma(U-U_0)}}. \tag{2.4}$$

This logistic function transforms U, whose elements are always positive, to a matrix M_{cs}, whose elements have values between zero and one. The value of scaling parameter γ controls steepness of a sigmoidal curve and the midpoint U_0 determines where 0.5 takes place in the map in Eq. (2.4). The map M_{cs} may be considered as a Waddington landscape of the high-dimensional gene expression data projected into a two-dimension plane. The basins of attractions of the CSM correspond to individual cellular states in the data.

Identify basins of cellular state map

In this process of identifying the basins of the CSM, all local minima in M_{cs} are searched first, leading to a pool of the minima in an increasing order. To construct the basin of the smallest local minimum (W), we first find the smallest local maximum, whose value is denoted as W_m, around this local minimum (W). Next we construct contours in the CSM that contains this minimum. The largest such contour value that is still smaller than W_m is the contour that contain the basin of this smallest local minimum (W). This searching procedure is then repeated for the second smallest minimum, and the rest of other minima. (More details can be found in Section B.1 in Appendices: B Additional file for Chapter 2).

Identify transition paths

Cellular state transition paths from one cellular state to the other are traced based on the CSM (M_{cs}). All cells in the first stage during transition processes need to be known in advance, which is the case for many temporal data. After locating the basins in the M_{cs}, for the cellular states at the first stage, we then identify its adjacent basins. The neighboring basin that has the smallest height of the barrier is locations of the cells for the next transition state, and then here it means the cells in the basin are at the second stage. If more than one of barriers have the similar heights, indicating a branch process takes place during transitions from the first stage to the second stage, we consider multiple cellular states emerge at the second stage. The procedure consisting of searching for adjacent basins, estimating heights of barriers, and identifying branching processes for each basin continues until all basins are analyzed. At the end of this procedure, the transition paths are also identified (Figure 2.1BC).

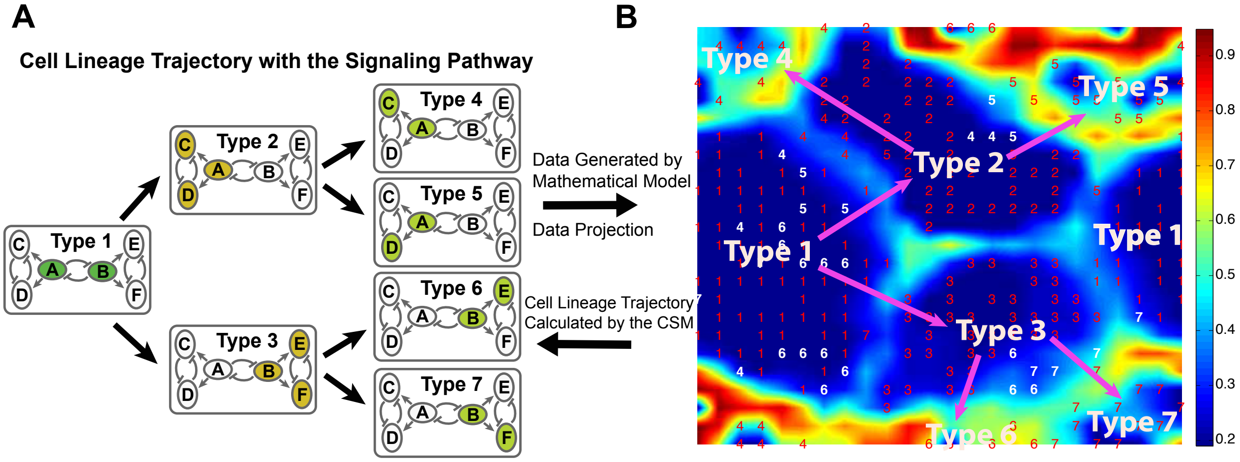


Figure 2.2: The CSM and cellular state transition paths based on the simulated model. (A) A three-stage lineage system. Stage 1 contains one type of cells in which the activated genes, A and B are highlighted in green; Stage 2 contains Type 2 cells and Type 3 cells. The activated genes, A, C, and D are highlighted in orange in Type 2 cells while the activated genes, B, E, and F are highlighted in orange in Type 3 cells. Stage 4 contains four types of cells: Type 4 cells, Type 5 cells, Type 6 cells, and Type 7 cells. The activated genes, A and C, A and D, B and E, or B and F are highlighted in light green in Type 4, Type 5, Type 6, and Type 7 cells, respectively. (B) The CSM with $N_g = 576(24 \times 24)$ grids is computed for the data of $N = 353$ single cells using $U_0 = 1.5$ and $\gamma = 1$. A red or white number shown in the CSM is a temporal stage of its corresponding cell in the data. A white number means its corresponding cell locates in an incorrect basin. A pink arrow shows a direction of a transition path.

Key parameters in SOMSC

In the standard SOM, a two-dimensional U-matrix may have the same size or different sizes in those two dimensions. To avoid bias on a particular gene or a subgroup of genes when applying the SOM to the single-cell data, here we consider both dimensions of a U-matrix to be the same .The total number of grid points in the CSM corresponding to the U-matrix is defined as $N_g = N_r \times N_r$ where N_r is the number of grids in each dimension of the CSM. The choice of N_g depends on the number of samples (e.g. the number of single cells), N, in order to compute the U-matrix more accurately. Naturally, the size of a U-matrix is proportional to the number of samples, such as $N_g = \beta N$, where β is a constant. Secondly, in the simulation N_g needs to be adjusted to avoid producing too many basins in a CSM, such as the case in which every one or two cells grouped as one basin. Two other key parameters are γ and U_0 in a CSM. As shown in the later sections, a CSM seems to produce the most consistent results when the choices of these two parameters enable a larger range of values of elements in M_{cs} from zero to one, allowing better separation between basins of cellular states.

2.2.4 Generate the simulation data

In order to effectively evaluate performance and choices of parameters of the SOMSC, we next construct a toy system consisting of a small number of genes to mimic single-cell gene expression data. There are three stages in the system, and in each stage one type of cells makes a transition to two other types of cells (Figure 2.2A). Together, seven types of cells with three branches present in the system. The cellular types are defined by the specific patterns of expression levels of the six genes (Figure 2.2A). Specifically, in Type 1 cells Gene A and Gene B are activated and all other four genes are silenced; in Type 2 cells Gene A, Gene C, and Gene D are activated; in Type 3 cells Gene B, Gene E, and Gene F are activated; when one of Gene A and Gene B and one of Gene C, Gene D, Gene E and Gene F are activated, four other types of cells in the third stage are then defined as Type 4, Type 5, Type 6, and Type 7 cells, respectively.

The system of three-toggle modules consisting of six genes is modeled through a system of stochastic differential equations [19, 47, 106]. Starting with only Type 1 cells in the system (i.e. the initial state), the expression values of each gene are then collected at three different temporal stages for each stochastic simulation: the early, the middle, and the final stage, in order to mimic a typical set of temporal single-cell data (See Section B.2 in Appendices: B Additional file for Chapter 2). Repeating the stochastic simulations using the same set of parameters and the same initial values of genes for 400 times produces a set of gene expression values, corresponding to 1200 sets of single-cell data.

2.3 Results

2.3.1 SOMSC on the simulation data

To mimic a typical size of experimental data, we randomly select expression levels of 353 cells out of the ones of 1200 cells collected in the simulation data. In the CSM calculated using the SOMSC, each cell is marked by its temporal state collected (Figure 2.2B). By tracking basins and analyzing heights of barriers, we obtain different cell types and their transition relationship (Figure 2.2B). Interestingly, in this case the adjacent basins of the basin of Type 1 cells contain all other types of cells from Type 2 to Type 7. However, the barriers between the basin of Type 1 cells and the basins of Type 4, 5, 6, and 7 cells are higher than those for the basins of Type 2 and Type 3 cells, suggesting two possible transition paths: one transition from Type 1 cells to Type 2 cells and the other from Type 1 cells to Type 3 cells (Figure 2.2B). Next, the barriers between the basin of Type 2 and those of Type 4 and

Type 5 are found to be lower than the ones for basins of Type 6 and Type 7 cells. So Type 2 cells make a transition to Type 4 cells or Type 5 cells. The barriers between basins of Type 3 cells and those of Type 6 and Type 7 cells have similar heights, indicating the next transition state of Type 3 cells is either Type 6 cells or Type 7 cells.

To study effects of the number of grids N_g on performance of the SOMSC, we systematically vary N_g and the number of observations N in the toy model (See Figure B.4). First we fix $N = 100$ observations (or cells) from the toy model but explore five different N_g (See Figure B.4A to B.4E). When N_g is too small (See Figure B.4AB) the CSM is unable to capture all the basins in the system whereas when N_g is too large (See Figure B.4E) the CSM tends to overpopulate the basins by grouping every one or two cells into one basin. It is found that the CSM profile becomes more consistent and reliable when N_g is in its middle range of values (See Figure B.4C and B.4D). Such trend remains when the number of observations (or cells) increases to $N = 200$ (See Figure B.4F to B.4I), and to $N = 353$ (See Figure B.4K to B.4O). Together, when β, the ratio between N_g over N, is in a range of $[1, 10]$, the patterns of basins and transition paths in the CSM start to become more consistent. In other words, given the number of observations, the size of the map in the SOMSC N_g needs to be explored until a ”convergent” pattern is observed.

It is observed that around 5% of the 353 cells are placed in the incorrect basins in the CSM (marked in white in Figure 2.2B). Such inconsistency might be due to noise in the data or choices of parameters in the SOMSC. Interestingly, if the data set is analyzed without involving the gene expression levels of those incorrect cells, the new CSM has no cells locating incorrect basins (See Figure B.5 and B.6 in Appendices: B Additional file for Chapter 2), suggesting that either those cells are less consistent compared to the rest of cells in the original data set or the SOMSC is too sensitive to the gene expression levels of those cells. Two other important parameters in determining the CSM are the midpoint of the logistic function (i.e. U_0) and the scaling factor (i.e. γ) in Eq. (2.4). We systematically explore

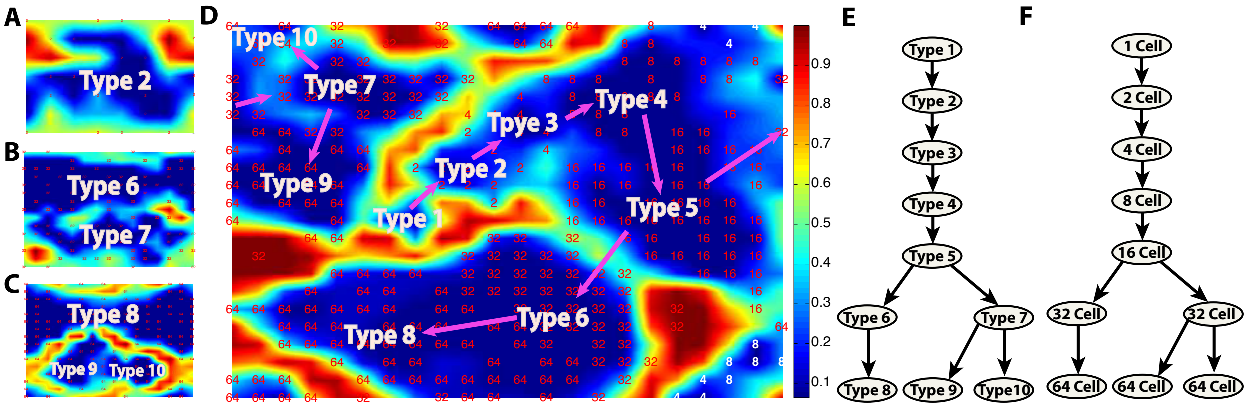


Figure 2.3: CSMs and a lineage trajectory are constructed using the qPCR data of mouse stem cells from zygote to blastocyst [46]. (A - C) CSMs obtained using data only at the second, sixth and seventh stages, respectively. A red or white number in (A, B, C) represents an index of stages when the expression levels of cells were measured. (A) Type 2 labels the only basin of cells in the CSM computed using the data only from the second stage. Here $N_g = 36$, $U_0 = 5$ and $\gamma = 0.01$. (B) Type 6 and Type 7 label two separate basins of the CSM computed using the data only from the sixth stage. Here $N_g = 196$, $U_0 = 2$ and $\gamma = 0.3$. (C) Type 8, Type 9, and Type 10 label three separate basins of the CSM using the data only from the seventh stage. Here $N_g = 196$, $U_0 = 2$ and $\gamma = 0.3$. (D) The CSM is computed using the data collected all seven stages with a total of $N = 442$ cells. Here $N_g = 484$, $U_0 = 2$ and $\gamma = 2$. Ten basins are labeled by Type 1, Type 2, . . . , and Type 10. A white number means its corresponding cell is located in an incorrect basin. A pink arrow indicates a direction of a transition path. (E) The state transition paths are derived from the CSM in (D). (F) The differentiation lineage tree of early mouse development was obtained in a previous study [46].

different values of those two parameters and their effects on the CSMs and the transition paths. The sigmoid's midpoint U_0 determines the range of the values of elements in M_{cs}. A larger value of U_0 usually leads to smaller values of elements of M_{cs} (e.g. most of elements in M_{cs} become smaller than 0.5 and some of them are close to zero) while a smaller value of U_0 leads to larger values of elements in M_{cs} (e.g. larger than 0.5 and close to one). For the scaling factor, a larger value of γ usually makes M_{cs} better cover the entire range of $[0, 1]$, however, sometimes it also makes many elements of M_{cs} close to 0 or 1. It is found that when the elements of M_{cs} are more evenly distributed in $[0, 1]$ by adjusting the parameters U_0, and γ, the computed CSM becomes more consistent and reliable (See Figure B.7 and B.8 in Appendices: B Additional file for Chapter 2).

2.3.2 SOMSC on experimental data

qPCR data of mouse embryo development from zygote to blastocyst

Previously, the expression levels of 48 genes at seven time points were measured using qPCR for mouse early embryonic development from zygote to blastocyst [46]. The raw data of the 442 single cells were normalized cell-wisely by the mean expression levels of two genes: Actb and Gapdh [46].

Two different approaches might be applied to such data set by either using the data at each temporal point individually or lumping the data of all seven stages into one set. For example, applying the SOMSC to the data at the second stage results in a CSM with one cell type (Figure 2.3A), and using the data point at the sixth stage or the seventh stage results in two cell types (Figure 2.3B) or three cell types (Figure 2.3C), respectively. However, such approach is unable to determine potential transition paths among cell types inherited in the data because different basins or cellular states are obtained using different CSMs.

Using all 442 cells collected at the seven stages simultaneously produces one CSM containing 10 basins (Figure 2.3D), and the relationship of those basins can then be analyzed to study state transitions. The basin labeled as a Type 1 cell is chosen based on those cells marked at the initial stage in the collected data [46]. The other nine basins are labeled by Type 2, . . . , Type 10. In the CSM, the Type 1 cell has three neighboring basins, and the barrier between the basin of the Type 1 cell and the basin of the Type 2 cell is found to be lower than those barriers separating with other basins, indicating the Type 1 cell makes a transition to the Type 2 cell. Similar analysis suggests that the Type 3 cell is the next transition state of the Type 2 since the corresponding barrier height is lower than others.

As seen in the CSM, clearly there is a transition from the Type 3 cell to the Type 4 cell. The height of the barrier between the basin of the Type 5 cell and the basin of the Type

4 cell is lower than others, showing that the Type 4 cell makes a transition to the Type 5 cell. The next transition states of the Type 5 cell are the Type 6 cell or the Type 7 cell because the heights of the barriers between them are lower than others, suggesting a branch process takes place. The barrier between the Type 8 cell and the Type 6 cell is rather low, indicating that the Type 6 cell becomes the Type 8 cell. Finally, two basins adjacent to the Type 7 cell have barriers of similar heights, indicating that there are two transitions from the Type 7 cell to the Type 9 cell or the Type 10 cell. As a result, seven stages containing two branches are identified, corresponding to the seven developmental stages [46]: 1-cell stage, 2-cell stage, . . . , 64-cell stage. Two major cell types (TE and ICM) arise at the 32-cell stage, and later the ICM cells differentiate to EPI or PE cells at the 64-cell stage (Figure 2.3E). To investigate each individual cell, one can index each cell by a proper order to scrutinize its location in the CSM for its transition capabilities or other properties relative to some other cells (see Figure B.9 in Appendices: B Additional file for Chapter 2).

It is not surprising that a very small number of cells (around 5% out of 442 cells marked in white color) that were collected at one developmental stage in the experiment are not exactly located in the corresponding basins of the CSM (Figure 2.3D). Interestingly, the ”mismatch” cells are found to be mostly collected in the 8-cell stage. Noise in the measurements, the small number of observations, and the choices of parameters used in the SOMSC may all contribute to this mismatch. To further study this, we next vary the sizes of mappings from $N_g = 484$ to $N_g = 900$, and find that the overall patterns of the lineage trees hardly change (See Figure B.10 in Appendices: B Additional file for Chapter 2). However, when we use $N_g = 100$ or $N_g = 3600$, the number of basins and the obtained transition paths start to become inconsistent (See Figure B.11 in Appendices: B Additional file for Chapter 2). Overall, it is important to vary the parameters used in the SOMSC in order to capture a reliable CSM with consistent cell types and transition paths using the noisy single-cell data.

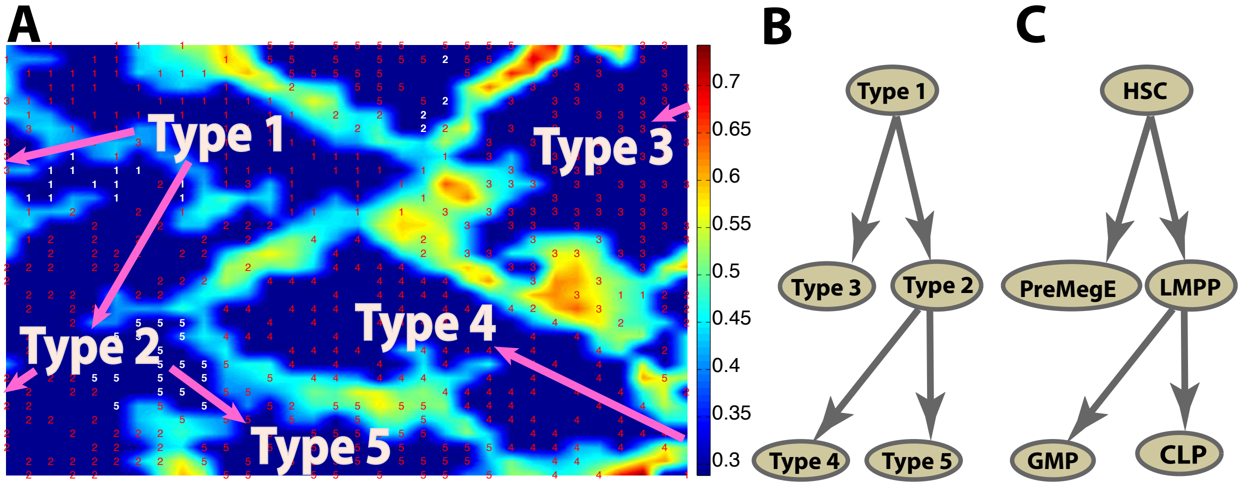


Figure 2.4: The CSM and a cell lineage trajectory are constructed using the qPCR data of mouse haematopoietic stem cells [99]. (A) The CSM is computed using $N_g = 1024$ based on $N = 597$ cells. Here $U_0 = 1.5$ and $\gamma = 0.88$. A red or white number represents a cell with a specified type given in the single-cell measurement [99]. The cells marked in white numbers are those in incorrect basins. A pink arrow is a direction of a transition path. (B) The state transition paths are obtained from the CSM in (A). (C) The lineage tree of mouse haematopoietic stem cells was obtained in the previous study [99].

qPCR data of mouse haematopoietic stem cells

In a previous study the expression levels of 24 genes including 18 core transcription factors were measured using qPCR for 597 mouse haematopoietic and progenitor stem cells [99]. The data were then normalized to the mean expression levels of two genes: Ubc and Polr2a [99]. After applying the SOMSC to this data set, we observe five different basins, indicating five possible cellular states inherited in the data marked by Type 1, Type 2, $\cdots$, Type 5 (Figure 2.4A). The Type 1 cell is identified using the prior knowledge given in the data [99]. Comparing all barriers surrounding the Type 1 cell, the height of barriers for Type 2 and Type 3 are much lower than the others. However, the height of the barrier for the Type 2 cell and the Type3 cell is similar, suggesting that the Type 1 cell may become either the Type 2 cell or the Type 3 cell. Similarly, it is found that the Type 2 cell may make a transition to either the Type 4 cell or the Type 5 cell.

Once the transition paths of the five types of cells are obtained (Figure 2.4B), we can easily establish a map between the transition paths and the well-known lineage trajectory of five mouse haematopoietic cell types [99]: haematopoietic stem cell (HSC), lymphoid-primed multipotent progenitor (LMPP), megakaryocyte-erythroid progenitor (PreMegE), common lymphoid progenitor (CLP) and graulocyte-monocyte progenitor (GMP) (Figure 2.4C).

Similar to the previous cases, a very small portion of cells fall into the incorrect basins (Figure 2.4A and Figure B.12 in Appendices: B Additional file for Chapter 2). For example, a small number of HSC cells (marked by white numbers in Figure 2.4A) are found located in the basin of the LMPP cells whereas a small number of CLP cells (also labeled in white) are found in the basin of LMPP cells. Missing entries in the raw data, the pre-processing method [99], the fact that LMPP is the intermediate cell types during transitions, and our choices of parameters in the SOMSC may all contribute to the mismatch. Also, similar to the study on the toy model, the choice of proper N_g is important in tracking the transition paths, and too small or too large values of N_g lead to inconsistent patterns of the CSMs (See Figure B.13 in Appendices: B Additional file for Chapter 2).

RNA-seq of human preimplantation embryos

In a previous single-cell RNA-seq analysis on human preimplantation embryos, 90 individual cells were sorted at seven stages: metaphse II oocyte, zygote, 2-cell, 4-cell, 8-cell, morula and late blastocyst, with two or three embryos used at each stage [166]. In this study, over 20,000 genes were measured using RNA-seq. Because the number of cells is small and the number of genes is very large in the data set, we only select those genes that are significantly expressed at least at one stage, leading to a system of 2,389 genes and 90 cells.

A CSM calculated by the SOMSC contains seven basins of cells (Figure 2.5A). A Type 1 cell is identified based on those cells in the metaphase II oocyte [166]. The rest of basins are then labeled by Type 2, Type 3, . . . , Type 7. The barrier between the Type 1 cell

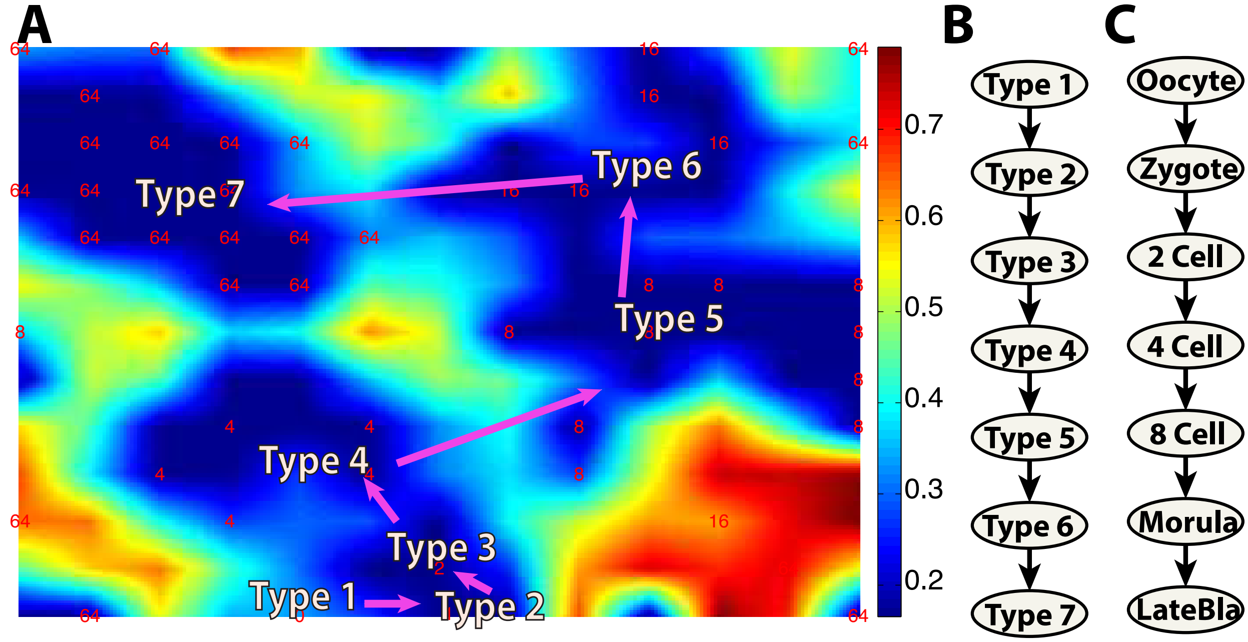


Figure 2.5: The CSM and lineage relationship are constructed using the RNA-seq data of human preimplantation embryonic cells from oocyte to late blastocyst [166]. (A) The CSM is calculated using $N_g = 169$ based on all $N = 90$ cells collected at differentiation stages. Here $U_0 = 20$ and $\gamma = 0.1$. A red or white number represents a stage of a cell measured. A white number means its corresponding cell is located in an incorrect basin. A pink arrow is a direction of a transition path. (B) The paths of transition are calculated from the CSM in (A). (C) The differentiation lineage tree of human preimplantation embryonic cells was obtained in the previous study [39]

and the Type 2 cell is found lower than those for the Type 3 cell, and the Type 7 cell. It indicates that the Type 1 cell make a transition to the Type 2 cell. Comparing the heights of barriers among the adjacent basins, the Type 2 cell likely make a transition to the Type 3 cell, and the next transition state of the Type 3 cell is the Type 4 cell that can make a transition to the Type 5 cell. Similar analysis shows that the Type 5 cell becomes the Type 6 cell that makes a transition to the Type 7 cell (Figure 2.5B). The observed cellular states and transition paths are consistent with the previous study (Figure 2.5C) [166]. The location of each cell and the distribution of cells in the CSM potentially provide additional information (e.g. signature genes for specific cellular types) for the lineage tree (See Figure B.14 in Appendices: B Additional file for Chapter 2).

It is found that too small or too large N_g in the SOMSC may result in inconsistent patterns of basins and transition paths in the CSMs (See Figure B.15 in Appendices: B Additional file for Chapter 2). However, by tuning the parameters in a systematic way, the SOMSC is able to obtain a "convergent" CSM and transition patterns.

RNA-seq of human skeletal muscle myoblasts

In a previous study single-cell RNA-seq of 271 cells collected from differentiating human skeletal muscle myoblasts (HSMM) were measured at 0, 24, 48 and 72h after switching human myoblasts to low serum [147]. 518 genes that were significantly and differently expressed across different time points and considered to be associated with myoblast differentiation were measured [147].

In the CSM consisting of seven basins marked by Type 1, Type 2, $\cdots$, Type 8 (Figure 2.6A), The Type 1 cell and the Type 2 cell were collected at 0h [147]. Analysis on the heights of barriers shows that a transition takes place from the Type 2 cell to the Type 3 cell, which can becomes the Type 4 cell. There are two adjacent basins next to the Type 4 cell, which

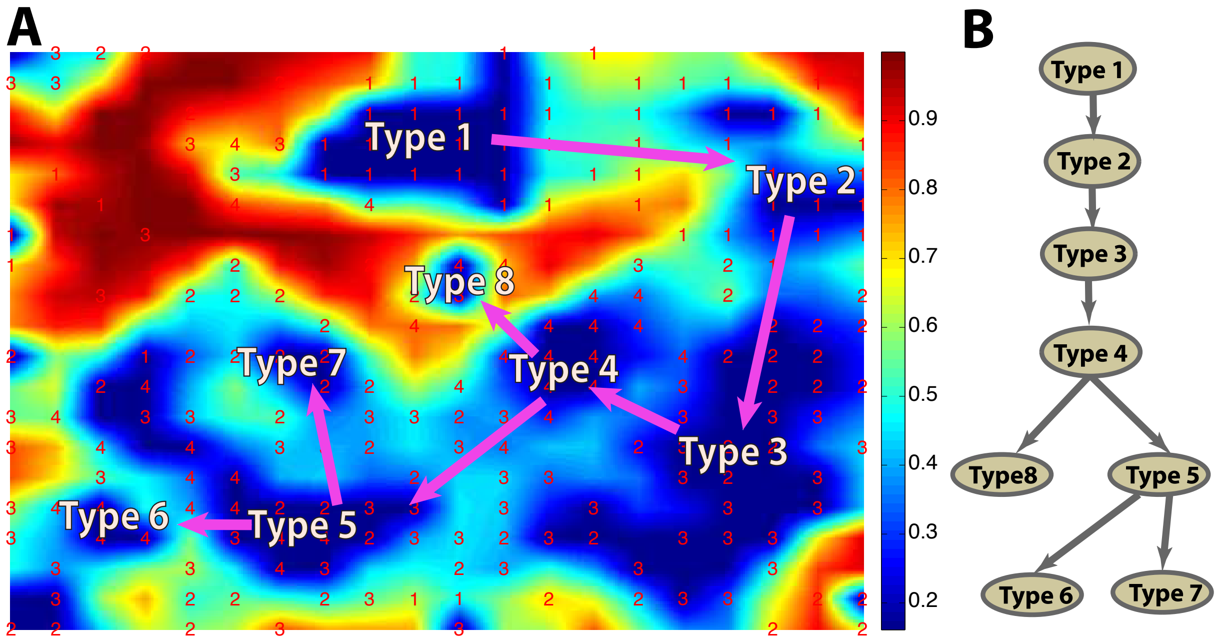


Figure 2.6: The CSM and lineage transition relationship are constructed using the single-cell RNA-seq data from human skeletal muscle myoblasts [147]. (A) The CSM is calculated using $N_g = 400$ based on $N = 271$ human skeletal muscle myoblasts cells collected at 0h, 24h, 48h, and 72h. Here $U_0 = 5$ and $\gamma = 0.8$. A red number is an ordered time point when the expression levels of cells were measured. The pink arrow is the direction of the transition path. (B) The lineage tree is predicted based on the CSM in (A).

may make a transition to the Type 8 cell or to the Type 5 cell. Finally, the Type 5 cell can become either the Type 6 cell or the Type 7 cell. The transition paths in a form of a lineage tree are then constructed accordingly (Figure 2.6B).

By comparing the temporal stage marked on each cell and the cell types identified using the SOMSC, we find that the transitions predicted from Type1, along Type 2, and Type 3, to Type 4 is consistent with the temporal sequence shown in the data. The CSM also predicts two different types of cells at 0h: Type 1 and Type 2, indicating a mixture of two subpopulations of cells at 0h. In addition, Type 3 consists of cells collected at both 24h and 48h. The CSM shows two branching processes taking place from the Type 4 cell to the Type 5 cell or to the Type 8 cell, and from the Type 5 cell to the Type 6 cell or to the Type 7 cell. The two branches are similar to those obtained by other algorithms [64, 147]. It is interesting to note that there are four types of cells collected at 24h, three types of cells collected at 48h, and three types of cells collected at 72h. These mixtures of different types of cells in multiple temporal stages suggest the gene expression plasticity might take place between the time points of measurements. Together, our simulations show capabilities of the SOMSC in predicting multiple cellular states and potential plasticity of subpopulations of cells.

2.4 Conclusion and Discussion

In this paper we have presented a self-organization-map based method for analyzing single-cell gene expression data that may contain multiple cellular states with transitions among them. Applications of the SOMSC to a set of simulated data and four sets of differentiation data have demonstrated strong capabilities and effectiveness of the SOMSC in identifying cellular states and their transitions.

A cellular state map (CSM) based on a U-matrix calculated from the SOM provides a global landscape view of cell differentiation or cellular state transitions. By estimating the heights of barriers between basins in a CSM, transition paths among the states are then identified. The location of each cell in the CSM may provide useful information on the cell's viability and potential of transitions to different cellular states. Such knowledge on individual cell in single-cell data is lacking in many other methods for single-cell analysis.

The major computational cost of the SOMSC comes from the iteration procedure in calculating the U-matrix in the SOM, with a complexity of $\mathcal{O}(N \cdot N_g \cdot D \cdot T)$ where D is the number of genes measured in the data, T is the number of iterations used in the SOM, and N is the number of samples in a single-cell data set[77]. In practice, D is usually around 1,000 (the number of genes significantly expressed), and both T and N are less than 1,000, implying a complexity of $\mathcal{O}(10^9)$ that the SOMSC is able to handle effectively.

Single-cell data are often used to identify cellular states in heterogeneous populations of cells [72]. However, the complexity in data visualization and analysis presents a major difficulty in distinguishing such subpopulations. The SOMSC may capture complex topological shapes in the data to identify those subpopulations due to the advantageous feature of the SOM unlike many other methods requiring convex or normal structure of the data [87]. Another major feature of the SOM is its capability of finding multiple minima as the entire space of feasible solutions in the SOM is searched until finding optimal solutions [87, 109]. This is consistent with the observations that the SOMSC is rather stable in searching for basins of attractions and transition paths in the CSM of single-cell data.

Several parameters in the SOMSC need to be tuned in order to obtain a reliable CSM. It is not surprising that given a number of samples (the number of cells and the number of genes measured), the number of grids for a U-matrix calculated by the SOM requires adjustment in order to obtain "convergence" of a corresponding CSM. The scaling parameter γ in Eq. (2.4) of a CSM was found to reduce noise effects in a U-matrix, allowing well-separated basins and well-defined barriers. Another important element to improve in the SOMSC is the approach in identifying basins and barriers. Matlab built-in contour construction method is currently used in this paper, and other algorithms may be further explored.

Noise and variability in single-cell data introduce another major complexity. In this work we have tried to reduce noise and variability effects by first removing those identified 'noisy' data from the training data sets. For example, in the case of the simulation data, cells located in incorrect basins are considered as the 'noisy' data. While a similar approach might be used for experimental data, identification of incorrect basins is clearly challenging, depending on availability of appropriate experimental measurements and prior knowledge on the systems. Potentially, machine-learning methods might be explored to enable reduction of noise effects for constructing a more consistent CSM. Other possibilities of improvement in this area include usage of different distance metrics (e.g. the diffusion metric [47]) instead of the standard Euclidean distance metric used in this work.

Previous works demonstrated that the confounding errors (e.g. batch errors) have great effects on single-cell data [14, 13]. PCA [117], surrogate variable analyses [78], probabilistic estimation of expression residuals [135, 136] or removal of unwanted variation [126] were explored to reduce such effects of confounders in gene expression measurements of the bulk cell populations [137]. Potentially, those methods could be extended to single-cell data. Other factors that are more unique to single-cell measurements, such as cell division, which may induce cell-cell variability, will provide an additional difficulty, for which a linear mixed model could be utilized [13]. In general, reducing the effects of confounding errors is essential to producing reliable classification of cellular states and identifying the transition paths among them.

A CSM produced by the SOMSC is similar to the gene expression landscape although a typical landscape is a function of each gene without dimension reduction. It would be interesting to make a comparison between a landscape computed by forward modeling based on a small size of network and a CSM generated by the SOMSC on single-cell data. Overall, the SOMSC provides a robust and convenient approach to classify the cellular states and to identify their transitions, and it is powerful in suggesting signature transcription factors, branching processes, and pseudo temporal orders of single cells.

Chapter 3

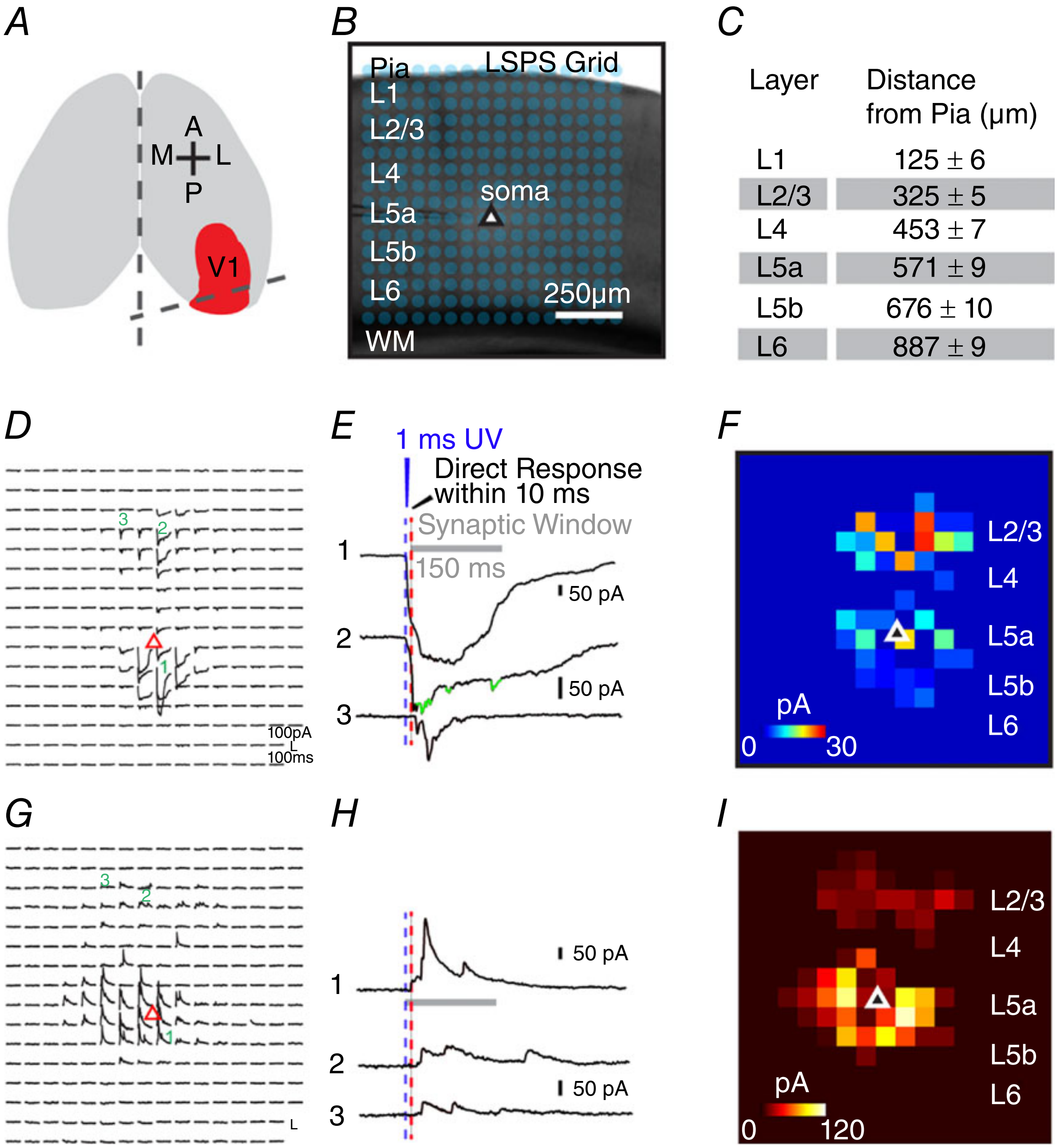
Infer synaptic connectivity in primary visual cortex

[Chapeter 3 is reprinted with the permission from Xiangmin Xu, Nicholas D. Olivas, Taruna Ikrar, Tao Peng, Todd C. Holmes, Qing Nie and Yulin Shi. Primary visual cortex shows laminar-specific and balanced circuit organization of excitatory and inhibitory synaptic connectivity. J Physiol 594.7 (2016)pp 1891-1910. ©2016 The Authors. The Journal of Physiology ©2016 The Physiological Society. [163]]

3.1 Introduction

The primary visual cortex (V1), similar to other cortical areas, contains excitatory and inhibitory cell types [94, 164]. Excitatory neurons are a principal cell group; they account for ∼80% of the whole cortical neuronal population and convey cortical excitation in both laminar and columnar dimensions. Given that cortical information processing is regulated by diverse types of inhibitory neurons (∼20% of the cortical neurons) and largely determined by local excitatory and inhibitory circuit interactions [52, 60, 48, 33], understanding how cortical circuits operate requires the clarification of both excitatory and inhibitory circuit connectivity. Although laminar organization of cortical circuits and the flow of cortical excitation in V1 have been established using anatomical and physiological methods [16, 31, 107], constructing layer-specific connectivity in cortical circuits based on identified neuronal types and their synaptic connections is much more difficult. Excitatory circuit connections to excitatory neurons in different V1 layers have been studied in vitro using physiological approaches such as paired intracellular recordings of synaptically connected neurons [145] and laser scanning photostimulation (LSPS) in which wider input sources are mapped to intracellularly recorded neurons [25, 168, 170]. The data derived have added important information on local functional circuit connections. However, the knowledge of intracortical synaptic connections to principal excitatory neurons in V1 still remains incomplete because most studies focus on excitatory neurons in a single cortical layer and there has yet to be a comprehensive and quantitative analysis that examines and compares excitatory synaptic connections to excitatory cell types across cortical layers 2/3-6. In addition, few studies have examined local laminar inhibitory connections to excitatory neurons in the sensory cortex [162, 68] and it remains unclear how their excitatory and inhibitory synaptic connections are spatially arranged on a layer-by-layer basis across local cortical circuitry. Although excitatory cells receive dense inhibitory neuronal innervation in highly localized microcircuits [37, 110], recent work suggests significant interlaminar or cross-laminar inhibitory connections to excitatory neurons [67, 133, 7, 108, 51, 119], prompting inhibitory cortical connections to be examined systematically using circuit mapping approaches.

In the present study, we used LSPS combined with whole cell recordings [162, 59, 75] to characterize and compare local circuit connectivity of the excitatory neurons in layers 2/3-6 in living mouse V1 slice preparations. We provide a quantitative assessment of the spatial distribution and input strength of excitatory and inhibitory connectivity with respect to individual pyramidal neurons across V1 laminar circuits, and construct laminar-specific



Layer	Distance from Pia (μm)
L1	125 ± 6
L2/3	325 ± 5
L4	453 ± 7
L5a	571 ± 9
L5b	676 ± 10
L6	887 ± 9

Figure 3.1: A, schematic V1 slice preparation: slices are made from mouse primary visual cortex, cut at a 75^o oblique angle relative to midline to preserve intracortical laminar connections. B, illustration of LSPS mapping of local cortical circuit input to single recorded cells. Excitatory neurons are recorded from binocular V1 region in whole cell mode, and the slice image is superimposed with a 16 × 16 LSPS mapping grid (blue dots, 65 μm^2 spacing) centred around the cell soma (triangle) and is aligned to the pial surface. Laminar boundaries are determined by cytoarchitectonic landmarks in bright-field slice images, validated by the boundaries determined by post hoc DAPI staining. C, average depth of laminar boundaries measured from the pial surface to the bottom edge of each layer ($n = 15$ slices). D, representative LSPS excitatory input map from voltage clamping an L5a pyramidal neuron at 70 mV in response to spatially restricted glutamate uncaging in the mapping grid (B). Each trace is plotted at the LSPS location shown in (B). E, detailed view of evoked EPSCs measured from the L5a pyramidal neuron at three respective locations numbered in (D). Trace 1 demonstrates a large 'direct response' resulting from uncaging at the perisomatic region. Trace 2 provides an example of a relatively small direct response in L2/3 from uncaging at the apical dendrite coupled with overriding synaptic inputs (shown in green). Trace 3 illustrates synaptic inputs (EPSCs) measured from a L2/3 location. Note the difference of amplitudes and latencies of direct and synaptic input responses, thus allowing for functional characterization. Empirically, responses within the 10 ms window from laser onset are considered direct, and exhibit a distinct shape (shorter rise time) and occurred immediately after glutamate uncaging (shorter latency). Synaptic events (i.e. EPSCs) are measured with the analysis window of $> 10 - 160$ ms after photostimulation (grey bar). For details, see Methods. F, colour-coded EPSC input map showing the overall spatial distribution and strength of excitatory inputs to the recorded L5a pyramidal cell. The map is constructed from the responses shown in (D); input responses per location are quantified in terms of average integrated EPSC strength within the analysis window, and colour coded according to the amplitude. G, representative LSPS inhibitory input map from voltage clamping an L5a pyramidal neuron at 5 mV in response to LSPS in the mapping grid similar to (D). H, examples of evoked IPSCs measured in an L5a pyramidal neuron at three respective locations numbered in (G). Trace 1 demonstrates large IPSCs measured near the cell soma. Traces 2 and 3 provide examples of interlaminar inhibition from L2/3. Consistent with excitatory inputs, IPSCs were measured with the analysis window of $> 10 - 150$ ms after photostimulation (grey bar). I, colour-coded IPSC input map showing the overall spatial distribution and strength of inhibitory inputs made to the L5a pyramidal cell.

synaptic wiring diagrams of excitatory neurons. The present study provides new knowledge on inhibitory laminar circuit connections and indicates that excitatory and inhibitory synaptic connectivity is spatially balanced onto V1 excitatory neurons.

3.2 Methods

3.2.1 Laminar circuit input analysis

Photostimulation can induce two major forms of uncaging responses: (1) direct glutamate uncaging responses (direct activation of the glutamate receptors of the recorded neuron) and (2) synaptically mediated responses (either EPSCs or IPSCs) resulting from the suprathreshold activation of presynaptic neurons. Uncaging responses within the 10 ms window from laser onset were considered direct, exhibited a distinct shape often with large amplitudes and occurred immediately after glutamate uncaging, as demonstrated in Fig. 3.1. Synaptic currents with such short latencies are not possible because they would have to occur before the generation of APs in photostimulated neurons. Therefore, direct responses need to be excluded from the synaptic input analysis. However, at some locations, synaptic responses were overriding on the relatively small direct responses and were identified and included in synaptic input analysis (Fig. 3.1). For data map analysis, we implemented a new approach for the detection and extraction of photostimulation-evoked postsynaptic current responses [132], which allows detailed quantitative analyses of both EPSCs and IPSCs (amplitudes and the numbers of events across LSPS stimulation sites). LSPS-evoked EPSCs and IPSCs were first quantified across the 16 × 16 mapping grid for each map and two to four individual maps were averaged per recorded cell, reducing the effect of spontaneous synaptic events. The analysis window (>10 ms to 160 ms) after photostimulation was chosen because photostimulated neurons fired most of their APs during this time (Fig. 3.2). Averaged

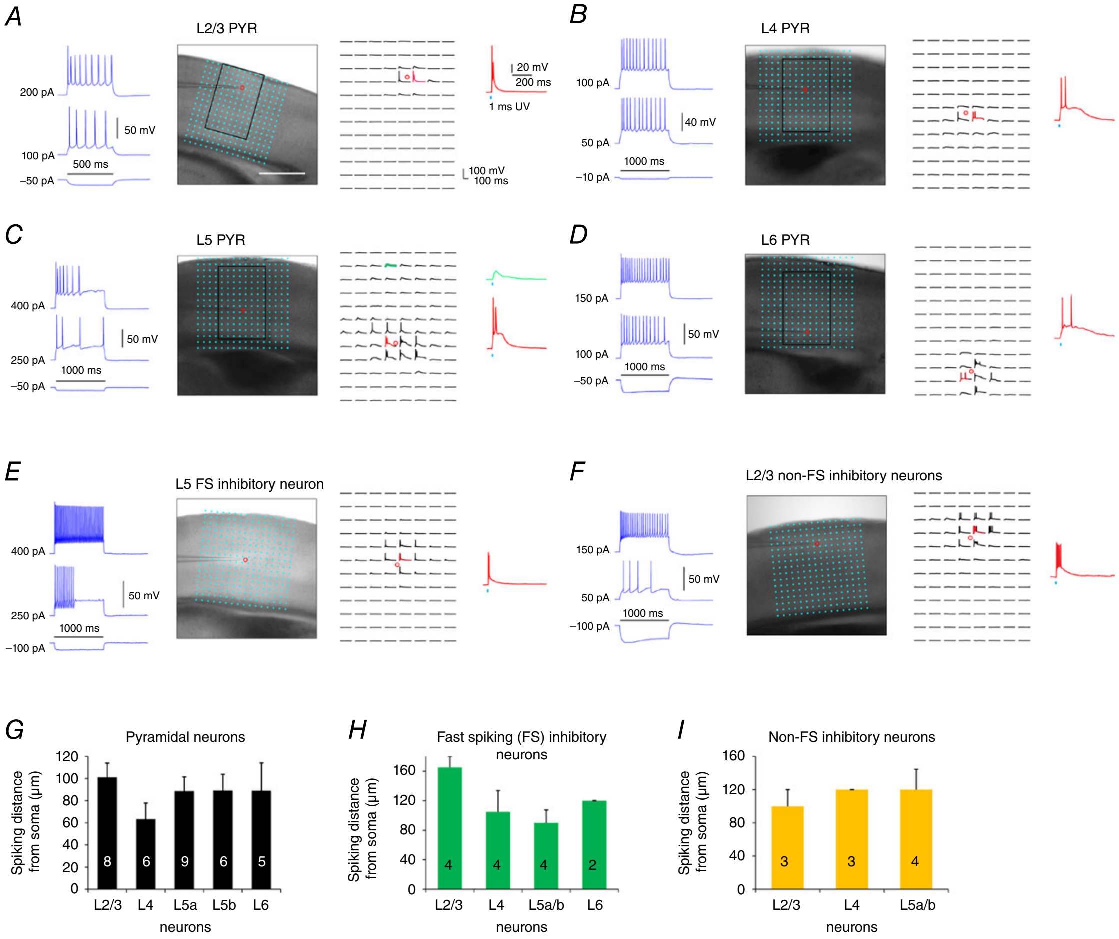


Figure 3.2: A-F, examples of photostimulation-evoked excitability profiles of pyramidal and inhibitory interneurons in different visual cortical layers. A, current injection responses of an example L2/3 pyramidal neuron are shown on the left; the image of V1 slice where the cell was recorded in layer 2/3 is superimposed with photostimulation sites (*, cyan dots, 65 μm^2 spacing) (middle) and the photostimulation responses of the recorded neuron are plotted at the beginning of stimulation onset (right). The individual responses are plotted relative to their spatial locations in the mapping array shown in the middle. The small red circle indicates the somatic location of the recorded neuron. One response trace with photostimulation-evoked APs is indicated in red, and shown separately by the side. Laser flashes (1 ms, 15 mW) were applied for photostimulation mapping. The scale in (A) is 500 μm. B-F, similarly formatted as in (A), with example L4, L5 and L6 pyramidal neurons, and L5 fast spiking inhibitory neurons and L2/3 non-fast spiking inhibitory neurons, respectively. C, two response traces with photostimulation-evoked subthreshold depolarization (green, photostimulation at the apical dendrite) and suprathreshold APs (red, perisomatic region) are shown separately. G-I, spatial resolution of LSPS evoked excitability of pyramidal neurons, fast-spiking and non-fast spiking inhibitory neurons was determined by measuring the LSPS evoked spike distance relative to soma location. Note that the spiking distance is measured as the 'vertical' distance (perpendicular to cortical layers) above and below the cell body. The numbers of recorded neurons are shown at the bar graphs. Data are presented as the mean ±SE.

maps were analysed and response measurements were assigned to individual laminar locations according to slice cytoarchitectonic landmarks and cortical depths from the pia surface (see below). Laminar distributions, average integrated input strength and the numbers of EPSCs measured in excitatory neurons were quantified. Input maps were plotted with average integrated EPSC or IPSC amplitudes, as well as evoked EPSC and IPSC numbers per location.

Because almost all layer 1 neurons are inhibitory cells, and pyramidal neurons with apical dendritic tufts in layer 1 could fire APs when their tufts were stimulated in layer 1 [25], EPSCs detected after photostimulation in layer 1 were not included in the analyses. However, because layer 1 neurons can provide inhibition to layer 2/3 neurons, we analysed IPSCs detected after photostimulation in layer 1.

3.2.2 Computational modeling

We adopted a discrete dynamical model [22] with the inference of the connectivity among excitatory neurons at four different cortical layers to describe and simulate photostimulation mapped circuit activities. The input data of the model were derived from the temporal data based on the LSPS-mapped synaptic inputs (EPSCs and IPSCs) to the representative excitatory neurons in different cortical layers. We extracted the temporal data from LSPS-mapped synaptic inputs to excitatory neurons, with six representative neurons from layer 2/3, four neurons from layer 4, seven neurons from layer 5 and four neurons from layer 6. The integration data of synaptic inputs was extracted at each 10 ms window using the detection method described above. To set a cut-off threshold for background noises, the data points (10 ms per point) with their strengths less than 50 pA/10 ms were set to zero. For photostimulation in each and specific cortical layer (i.e. L2/3, L4, L5 and L6), the overall activation strength was calculated by summing the area under each curve of temporal input evolution across all layers, and the relative laminar activation was obtained by comparing its area under the curve with the overall activation across layers.

Our model consists of the four cortical layers (L2/3, L4, L5 and L6) of excitatory neurons. In this model, a matrix, $W = (W_{ij})n$, n is used to represent the connectivity strength among different cortical layers. If $W_{ij} > 0$, layer j receives excitatory input from layer i. If $W_{ij} < 0$, layer j receives inhibitory input from layer i. If $W_{ij} = 0$, there is no direct connection between layers i and j. Specifically, for the present study, the $W = (W_{ij})_{4,4}$ has 16 entries, and the entries for L2/3 with L2/3, L4, L5 and L6 are W_{11}, W_{12}, W_{13} and W_{14}. The entries for L4 with L2/3, L4, L5 and L6 are W_{21}, W_{22}, W_{23} and W_{24}. The entries for L5 with L2/3, L4, L5 and L6 are W_{31}, W_{32}, W_{33} and W_{34}. The entries for L6 with L2/3, L4, L5 and L6 are W_{41}, W_{42}, W_{43} and W_{44}. The data from the mapping experiment were then used to fit the linear system to solve for W. The data fitting is further constrained by including the prior knowledge of the connectivity of cortical layers, as well as a term that controls the density of the connections, which may potentially remove very weak interlaminar connections. The model provides an optimal estimate for the connectivity strength matrix by minimizing the difference between the model-calculated signals and the measured experimental signals. Such an approach has been successfully used to obtain the gene regulatory network based on gene expression data [22].

Mathematically, the objective function for fitting the model to the data is:

$$W^* = \arg\min_W (g(W) + \alpha ||W||_1 + \beta ||W \circ W^0||_1)$$

where

$$g(W) = \sum_{h=1}^{p} \sum_{k=1}^{n} \sum_{l=1}^{m} ||W x_l^{k,h} - u_l^{k,h}||_2^2$$

display math In the objective function, α and β are the non-negative numbers and W^0

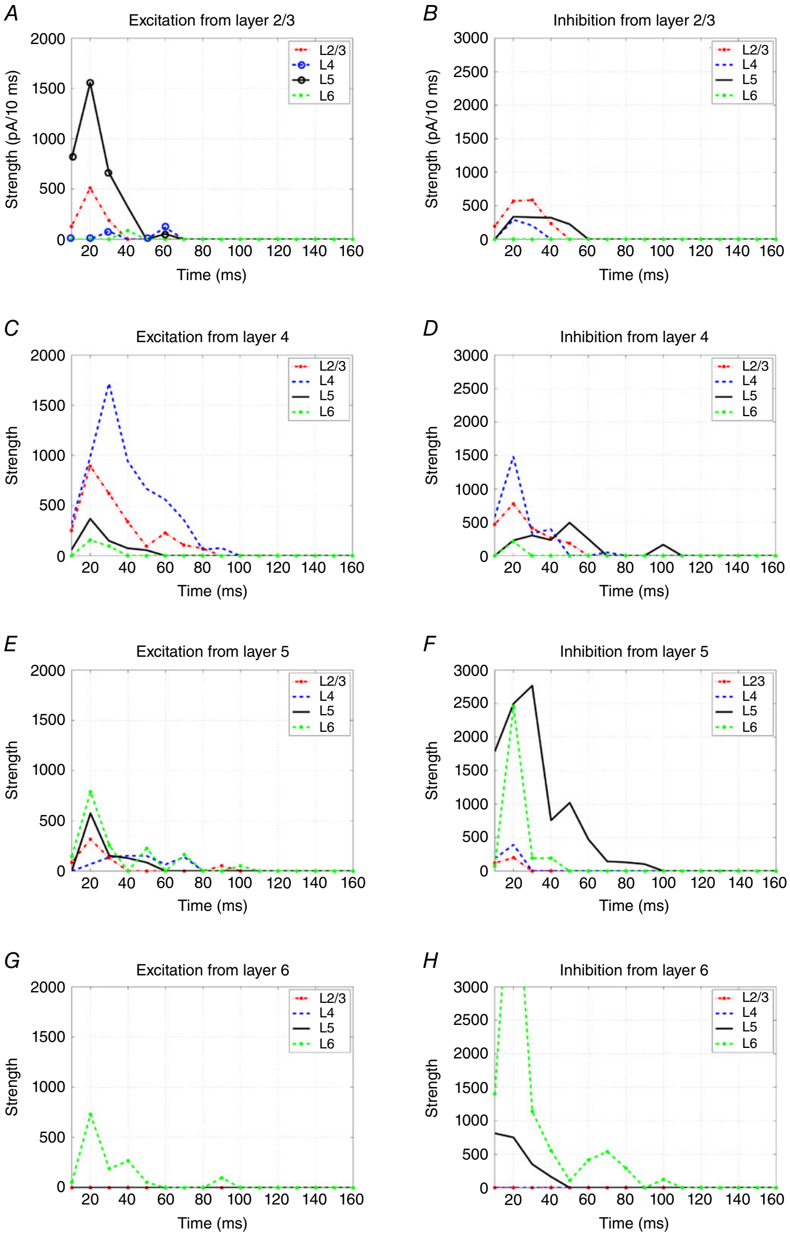


Figure 3.3: The temporal evolution and laminar distributions of local V1 circuit inputs to excitatory pyramidal neurons across different cortical layers in response to layer-specific photostimulation. A, C, E and G, each showing excitatory inputs to L2/3, L4, L5 and L6 excitatory neurons in response to photo- stimulation in L2/3, L4, L5 and L6, respectively. B, D, F and H, each showing inhibitory inputs to L2/3, L4, L5 and L6 excitatory neurons in response to photostimulation in L2/3, L4, L5 and L6, respectively. The x-axis represents the time (ms) and the y-axis represents the input strength (integrated synaptic input strength, pA/10 ms). Lines with different colours indicate the plots of inputs to the specified cortical layers.

is the prior knowledge of the connectivity strength of cortical layers (i.e. laminar relative activation; see above). The $m(=4)$ is the number of the given laminar photostimulation, the $n(=15)$ is the number of data pairs at each given photostimulation, the $p(=2)$ is the number of types of synaptic inputs (excitation and inhibition). If there is a connection from layer i to layer j, $w_{ij}^0 = 0$; otherwise, $w_{ij}^0 = 1$. $||W||_1$ is the term for controlling the sparseness of the network (i.e. the density of the connections), and $||W \circ W^0||_1$ is for incorporating the prior information where $\circ$ is the operation for the entry-wise multiplication. In the model, we used a simplified version of the network map as the prior information. The measured experimental data input, math formula, represents input strengths of L2/3, L4, L5 and L6 at time point k with a lth layer photostimulation ($l = 1, 2, 3, 4$; photostimulation in L2/3, L4, L5 or L6), given hth type laminar photostimulation (for measuring excitation or inhibition), and $u_l^{k,h}$ represents input strengths of L2/3, L4, L5 or L6 at time point math formula at the lth layer photostimulation given hth type laminar photostimulation as the corresponding output of $x_l^{k,h}$. $Wx_l^{k,h}$ is the model-calculated input strengths of L2/3, L4, L5 or L6 at time point $k+1$ at the lth given hth type laminar photostimulation. The objective of the modelling is to obtain the optimal connectivity strength matrix W^* by minimizing the difference between the model-calculated input strengths $Wx_l^{k,h}$ and the measured experimental input strengths $u_l^{k,h}$. For each given laminar photostimulation, there were 16 time points, including 15 data pairs ($x_l^{k,h}$,$u_l^{k,h}$), where $k = 1, 2, ..., 15$, $l = 1, 23, 4$ (photostimulation in L2/3, L4, L5 or L6 respectively) and $h = 1, 2$ (measuring excitation or inhibition). Altogether, 120 data pairs (15 × 4 × 2) were employed to train the model. We used the standard 10-fold cross-validation technique from machine learning to determine the values of α and β. Then, the optimal connectivity matrix W^* was calculated using the algorithm described in our previous study [22].

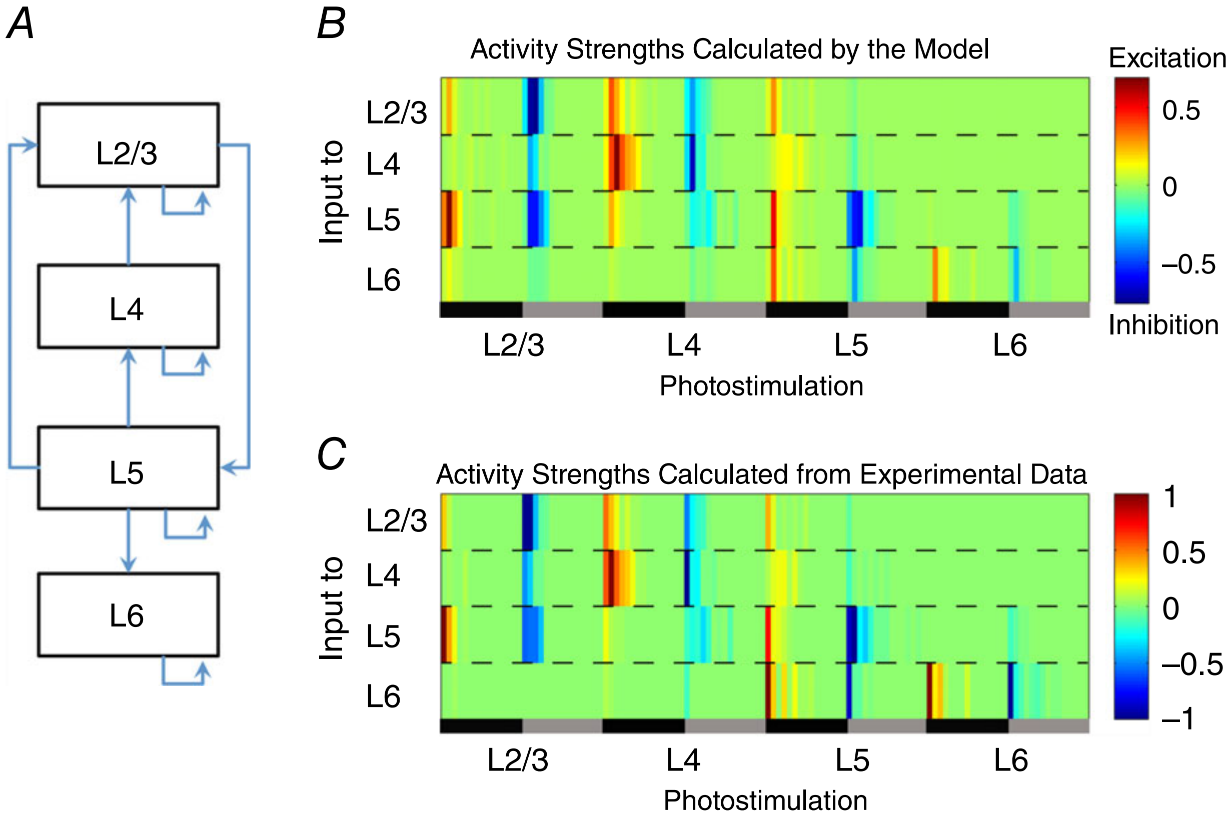


Figure 3.4: Photostimulation mapped circuit activities are simulated by the discrete dynamical model. A simplified laminar connectivity map (A) and the temporal evolution data across layers are used for the prior information in the model. B, synaptic input strengths at given time points in different cortical layers simulated by the discrete dynamical model with the optimal connectivity matrix (see Methods) [0.7818, -0.04, -0.0049, 0; 0, 0.6933, 0.0129, 0; 0.2200, 0, 0.6087, 0; 0, 0, 0.1086, 0.3108]. C, synaptic input strengths at given time points in different layers observed by experiments. The x-axis of (B) and (C) represents the conditions of photostimulation in L2/3, L4, L5 and L6. For each photostimulation in a specified layer, the dark zone indicates the temporal domain (150 ms post-photostimulation) of excitatory inputs evoked by photostimulation, whereas the grey zone indicates the temporal domain (150 ms post-photostimulation) of inhibitory inputs. The y-axis of (B) and (C) represents the input strengths for L2/3, L4, L5 and L6, according to the colour scales, in which the relative activation strengths are coded at given time points.

3.3 Results

we examined the temporal features of laminar distributions of local V1 circuit inputs to excitatory pyramidal neurons. To complement the static wiring, the temporal evolution of excitatory and inhibitory synaptic inputs to excitatory neurons in different cortical layers is shown in Fig. 3.3. Based on the data derived from a typical subset of sampled neurons, a great majority of inputs are observed to occur within 100 ms of layer-specific photostimulation, with the peak input strengths located between 20 and 40 ms. In Fig. 3.3A, the L5 excitation in response to L2/3 photostimulation exhibits a single peak that is temporally correlated with the L2/3 excitation, and the L2/3 → L5 input quickly falls off to the baseline at ∼50 ms post-photostimulation. If there were significant polysynaptic activation, the L5 excitation should have been broader with multiple peaks, and the L4 excitation would be much stronger. In Fig. 3.3C, L4 → L2/3 or L5 excitation peaks earlier than the L4 excitation. Their excitation falls rapidly, as predicted by the time course of the direct inputs. In addition, layer-specific inhibition generally decays quickly (Fig. 3.3B, D, F and H), matching the time course of layer-restricted excitation (Fig. 3.3A, C, E and G). Taken together, this temporal analysis supports the conclusion that LSPS maps direct synaptic inputs, and argues strongly against the possibility that feed-forward synaptically driven events could account for most of the input mapping responses measured.

A discrete dynamical model [22] was used to infer the connectivity among excitatory neurons at four different cortical layers. As shown in Fig. 3.4, the photostimulation-evoked circuit input activities are simulated well by the discrete dynamical model. The model simulations further support that the temporal evolution of excitatory and inhibitory synaptic inputs to excitatory neurons is laminar-specific and balanced in the visual cortex. The proof-of-principle demonstration indicates that our photostimulation-based experiments are capable of generating effective spatiotemporal data that can be directly used through computational modelling to predict the cortical circuit operations.

3.4 Conclusion

On the basis of the functional circuit mapping, the present study has provided important information for further computational modelling analysis. In particular, through in silico perturbation of circuit nodes in the model, dynamic network characteristics beyond the direct laminar circuit connections may be obtained. For example, an early initiating event in visual critical period plasticity is disinhibition in L2/3. One day of monocular deprivation during the critical period reduces excitatory drive onto parvalbumin-positive interneurons in binocular V1. This decrease in cortical inhibition is permissive for synaptic competition between excitatory inputs from each eye and is sufficient for subsequent shifts in excitatory neuronal ocular dominance [75]. Although the impact in L2/3 has been directly assessed, whether the effects of disinhibition may be expanded to other cortical layers remains to be explored. To address this and other related questions, we aim to test the circuit model in the future by blocking inhibitory projections from L2/3 inhibitory neurons to excitatory neurons. Reciprocally, physiological mapping experiments can be designed to test the predictions made by the model. Given that L2/3 neurons send strong projections to L5, we predict that the disinhibition effect would propagate to L5 for the laminar shift of cortical plasticity. Taken together, the interplay between modelling and experiment will probably provide new insights that could not be obtained by the experimental approach alone because the model drives the experimental design.

Chapter 4

Study Controlling Factors in Mouse Embryonic Epidermal Development

[Chapeter 4 is an ongoing project. Tao Peng, Xing Dai, Qing Nie. Study Controling Factors in Mouse Embryonic Epidermal Development.]

4.1 Background

Mammalian epidermis is a remarkable organ that must self-renew throughout life to maintain tissue homeostasis and repair because of the ability of self-renewal, proliferation and terminal differentiation of epidermal stem/progenitor cells [104, 105]. In the mouse, at around embryonic (E) day 9.5 the single-layered surface ectoderm begins to generate multiple subsequent lineages, such as the interfollicular epidermis, hair follicle, and sebaceous gland [74]. The biochemical hallmark keratin (K) 8/18 expression is switching off [74]. Meanwhile, K5 and K14, the marker of the future basal layer of mature epidermis is turning on. Starting at E14.5, the asymmetric division of proliferative basal cells results in the formation of a transient suprabasal layer, called intermediate cell layer and then at E15.5 the intermediate intermediate cells mature into spinous cells, which further differentiate into granular keratinocytes by E16.5. Finally the cornified layers are formed at E18.5 and they are essential for the organisms survival [74]. How embryonic epidermal morphogenesis is orchestrated at a molecular level to achieve the correct size and cell type proportions within the interfollicular epidermis remains poorly understood.

In this study, we investigate the transcriptional mechanisms governing growth and differentiation in developing epidermis. Previous work shows that many important genes or proteins are involved in epidermal proliferation and differentiation such as TGFα, TGFβ, Nothing Signaling, p63, cMyc, Zeb1 and Ovol gene circuitry [158]. TGFα is mainly expressed in the basal, proliferative layer of the skin epidermis and TGFβ presents in the suprabasal, differentiating layers [2, 88, 111, 139]. Notch signaling plays an important role in regulating the differentiation of basal cells as they are in the spinous layer [124]. p63 controls the stratification and proliferation of epidermis [12]. cMyc expression occurs in the basal cells and its constitutive overexpression in cultured keratinocytes result in progressively reduced growth, precocious terminal differentiation, and loss of cells [40]. Germline ablation of Ovol1 causes a thickened epidermis at birth with expanded spinous layers [24, 103]. Ovol2 represses keratinocyte transient proliferation and terminal differentiation through inhibiting c-Myc and Notch1 respectively [159]. Through the regulation relationship between the genes we construct the regulatory gene network which controls the cell lineage of cells in different layers, basal layer, spinous layer, and granular layer (Fig 1). In addition, cell-cell interaction plays an important role to regulate the cell proliferation and cell differentiation of cells at different stages. Especially, cell-cell communication such as Notch signaling and TGFβ entail representing receptors/ligands on the cell membrane surface with corresponding binding kinetics to ligands/receptors of adjacent cells.

This study is to integrate the mathematical models and experimental data to study the controlling factors of early embryonic epidermal development. In the mathematical model, we employed the gene regulatory network of regulators and cell lineage model at the first time. It provides an insightful way to explore the functions of the genes during the epidermal development.

4.2 Mathematical model of mouse embryonic epidermal development

TGFβ and Notch are the regulating factor in the cell niche and it means that they are the inputs of the signaling pathway. We use the following gene regulatory network model and the cell lineage model to study the dynamical development of mouse embryonic epidermis.

4.2.1 The mathematical model of the gene regulatory network

We model the multiscale mathematical modeling based on the following function.
X regulates Y positively,

$$f_+([X], K_{X-Y}, k_{X-Y}, n_{X-Y}) = \frac{k_{X-Y}[X]^{n_{X-Y}}}{K_{X-Y}^{n_{X-Y}} + [X]^{n_{X-Y}}} \tag{4.1}$$

X regulates Y negatively,

$$f_-([X], K_{X-Y}, k_{X-Y}, n_{X-Y}) = \frac{k_{X-Y}}{K_{X-Y}^{n_{X-Y}} + [X]^{n_{X-Y}}} \tag{4.2}$$

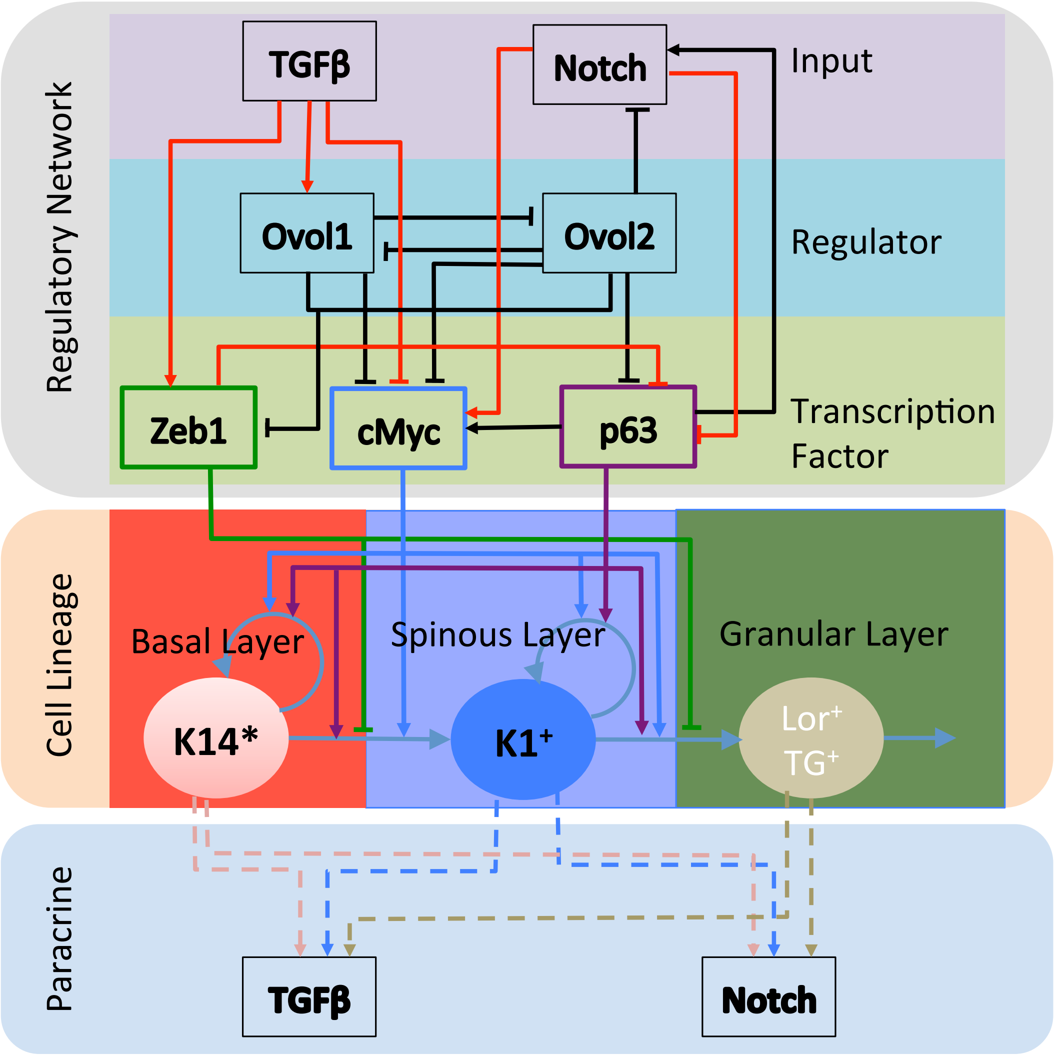


Figure 4.1: Critical morphological and molecular events during epidermal morphogenesis. The rectangles are the genes. The circles are different kinds of cells (basal cell, spinous cell, and granular cell). The solid lines between rectangles with arrows are positive regulations. The solid lines between rectangles with bars are inhibitive regulations.The dash lines with arrows are secreting processes of the three kinds of cells. The slide lines with arrows between cells are differentiation. The slide curve lines with arrows on cells are proliferation.Red solid lines are from references and the black lines are from microarray data.

$[X]$ is the concentration of X or the number of X cell. d_X is the degradation rate of X. The model of regulatory network

$$\begin{aligned}\frac{d[TGF\beta]}{dt} =& k_{K14-TGF\beta}[K14] + k_{K1-TGF\beta}[K1] + k_{K1TP-TGF\beta}[K1TP] \\ &+ k_{Lor-TGF\beta}[Lor] - d_{TGF\beta}[TGF\beta];\end{aligned} \tag{4.3}$$

$$\begin{aligned}\frac{d[Notch]}{dt} =& f_+([p63], K_{p63-Notch}, k_{p63-Notch}, n_{p63-Notch}) \\ & f_-([ovol2], K_{ovol2-Notch}, k_{ovol2-Notch}, n_{ovol2-Notch})[K14] \\ &+ k_{K1-p63-Notch} f_+([p63], K_{p63-Notch}, k_{p63-Notch}, n_{p63-Notch}) \\ & f_-([ovol2], K_{ovol2-Notch}, k_{ovol2-Notch}, n_{ovol2-Notch})[K1] \\ &+ k_{K1TP-p63-Notch} f_+([p63], K_{p63-Notch}, k_{p63-Notch}, n_{p63-Notch}) \\ & f_-([ovol2], K_{ovol2-Notch}, k_{ovol2-Notch}, n_{ovol2-Notch})[K1TP] \\ &+ k_{Lor-p63-Notch} f_+([p63], K_{p63-Notch}, k_{p63-Notch}, n_{p63-Notch}) \\ & f_-([ovol2], K_{ovol2-Notch}, k_{ovol2-Notch}, n_{ovol2-Notch})[Lor] - d_{Notch}[Notch];\end{aligned} \tag{4.4}$$

$$\begin{aligned}\frac{d[ovol1]}{dt} =& f_+([TGF\beta], K_{TGF\beta-ovol1}, k_{TGF\beta-ovol1}, n_{TGF\beta-ovol1}) \\ &+ f_-([ovol2], K_{ovol2-ovol1}, k_{ovol2-ovol1}, n_{ovol2-ovol1}) - d_{ovol1}[ovol1];\end{aligned} \tag{4.5}$$

$$\frac{d[ovol2]}{dt} = f_-([ovol1], K_{ovol1-ovol2}, k_{ovol1-ovol2}, n_{ovol1-ovol2}) - d_{ovo2}[ovol2][ovol1]; \tag{4.6}$$

$$\begin{aligned}\frac{d[zeb1]}{dt} =& f_+([TGF\beta], K_{TGF\beta-zeb1}, k_{TGF\beta-zeb1}, n_{TGF\beta-zeb1}) \\ &+ f_-([ovol1], K_{ovol1-zeb1}, k_{ovol1-zeb1}, n_{ovol1-zeb1}) \\ &+ f_-([ovol2], K_{ovol2-zeb1}, k_{ovol2-zeb1}, n_{ovol2-zeb1}) - d_{zeb1}[zeb1]; \end{aligned} \tag{4.7}$$

$$\begin{aligned}\frac{d[cMyc]}{dt} =& f_+([Notch], K_{Notch-cMyc}, k_{Notch-cMyc}, n_{Notch-cMyc}) \\ &+ f_+([p63], K_{p63-cMyc}, k_{p63-cMyc}, n_{p63-cMyc}) \\ &+ f_-([TGF\beta], K_{TGF\beta-cMyc}, k_{TGF\beta-cMyc}, n_{TGF\beta-cMyc}) \\ &+ f_-([ovol1], K_{ovol1-cMyc}, k_{ovol1-cMyc}, n_{ovol1-cMyc}) \\ &+ f_-([ovol2], K_{ovol2-cMyc}, k_{ovol2-cMyc}, n_{ovol2-cMyc}) - d_{zeb1}[zeb1]; \end{aligned} \tag{4.8}$$

$$\begin{aligned}\frac{d[p63]}{dt} =& f_-([Notch], K_{Notch-p63}, k_{Notch-p63}, n_{Notch-p63}) \\ &+ f_-([zeb1], K_{zeb1-p63}, k_{zeb1-p63}, n_{zeb1-p63}) \\ &+ f_-([ovol2], K_{ovol2-p63}, k_{ovol2-p63}, n_{ovol2-p63}) - d_{p63}[p63]; \end{aligned} \tag{4.9}$$

4.2.2 The mathematical model of the cell lineage

X regulates the self-renew of cell Y positively,

$$f_+([X], K_{X-Y-f}, k_{X-Y-f}, n_{X-Y-f}) = \frac{k_{X-Y-f}[X]^{n_{X-Y-f}}}{K_{X-Y-f}^{n_{X-Y-f}} + [X]^{n_{X-Y-f}}} \tag{4.10}$$

X regulates the proliferation of cell Y positively,

$$f_+([X], K_{X-Y-\mu}, k_{X-Y-\mu}, n_{X-Y-\mu}) = \frac{k_{X-Y-\mu}[X]^{n_{X-Y-\mu}}}{K_{X-Y-\mu}^{n_{X-Y-\mu}} + [X]^{n_{X-Y-\mu}}} \tag{4.11}$$

X regulates the self-renewal of cell Y negatively,

$$f_+([X], K_{X-Y-f}, k_{X-Y-f}, n_{X-Y-f}) = \frac{k_{X-Y-f}}{K_{X-Y-f}^{n_{X-Y-f}} + [X]^{n_{X-Y-f}}} \tag{4.12}$$

X regulates the proliferation of cell Y negatively,

$$f_+([X], K_{X-Y-\mu}, k_{X-Y-\mu}, n_{X-Y-\mu}) = \frac{k_{X-Y-\mu}}{K_{X-Y-\mu}^{n_{X-Y-\mu}} + [X]^{n_{X-Y-\mu}}} \tag{4.13}$$

$$\begin{aligned} f_{K14} =& f_+([cMyc], K_{cMyc-K14-f}, k_{cMyc-K14-f}, n_{cMyc-K14-f}) \\ &+ f_+([p63], K_{p63-K14-f}, k_{p63-K14-f}, n_{p63-K14-f}) \end{aligned} \tag{4.14}$$

$$\begin{aligned}\mu_{K14} =& f_{-}([zeb1], K_{zeb1-K14-\mu}, k_{zeb1-K14-\mu}, n_{zeb1-K14-\mu}) \\ &+ f_{+}([p63], K_{p63-K14-\mu}, k_{p63-K14-\mu}, n_{p63-K14-\mu}) \\ &+ f_{+}([cMyc], K_{cMyc-K14-\mu}, k_{cMyc-K14-\mu}, n_{cMyc-K14-\mu}) \\ f_{K1} =& f_{+}([cMyc], K_{cMyc-K1-f}, k_{cMyc-K1-f}, n_{cMyc-K1-f}) \\ &+ f_{+}([p63], K_{p63-K1-f}, k_{p63-K1-f}, n_{p63-K1-f})\end{aligned} \tag{4.15}$$

$$\begin{aligned}\mu_{K1} =& f_{-}([zeb1], K_{zeb1-K1-\mu}, k_{zeb1-K1-\mu}, n_{zeb1-K1-\mu}) \\ &+ f_{+}([p63], K_{p63-K1-\mu}, k_{p63-K1-\mu}, n_{p63-K1-\mu}) \\ &+ f_{+}([cMyc], K_{cMyc-K1-\mu}, k_{cMyc-K1-\mu}, n_{cMyc-K1-\mu})\end{aligned} \tag{4.16}$$

$$\begin{aligned}\mu_{K1TP} =& f_{-}([zeb1], K_{zeb1-K1TP-\mu}, k_{zeb1-K1TP-\mu}, n_{zeb1-K1TP-\mu}) \\ &+ f_{+}([Notch], K_{Notch-K1TP-\mu}, k_{Notch-K1TP-\mu}, n_{Notch-K1TP-\mu}) \\ &+ f_{+}([p63], K_{p63-K1TP-\mu}, k_{p63-K1TP-\mu}, n_{p63-K1TP-\mu}) \\ &+ f_{+}([cMyc], K_{cMyc-K1TP-\mu}, k_{cMyc-K1TP-\mu}, n_{cMyc-K1TP-\mu})\end{aligned} \tag{4.17}$$

$$\frac{[K14]}{dt} =(2f_{K14} - 1)\mu_{K14}[K14] \tag{4.18}$$

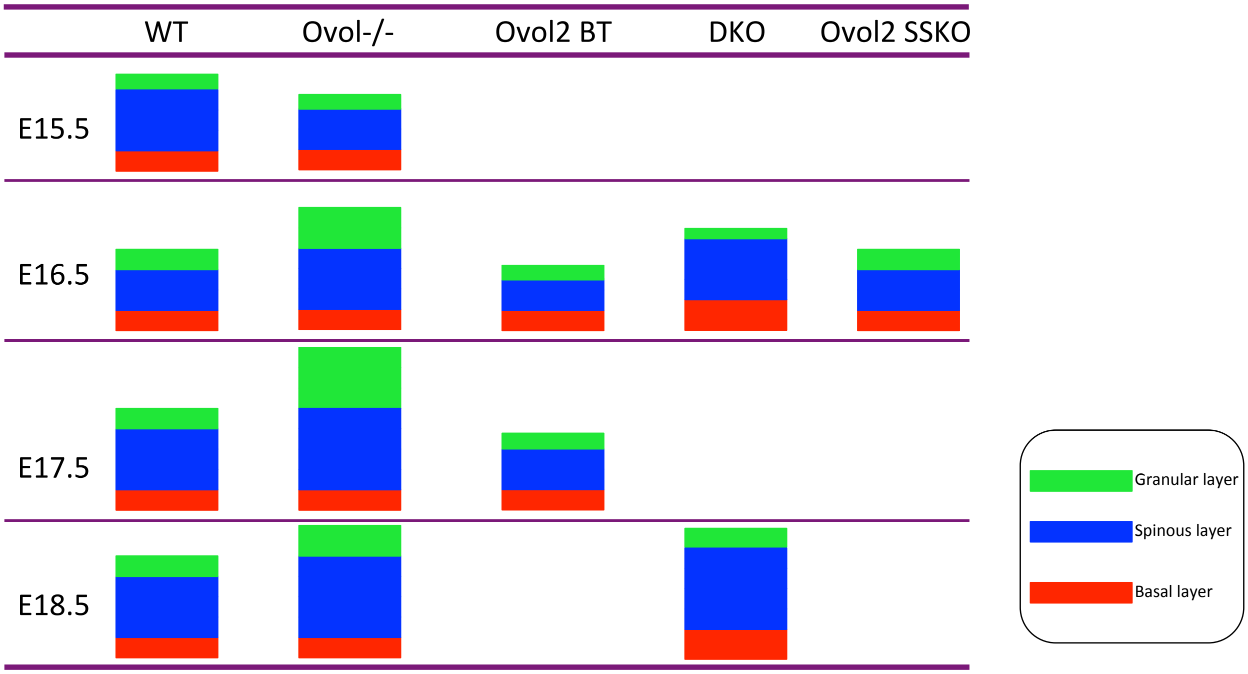


Figure 4.2: Skin epidermis thickness with three layers, basal layer, spinous layer, and granular layer at different stages E15.5, E16.5, E17.5, and E18.5 under different conditions, WT, Ovol1 knockout, Ovol2 overexpressed, double knockout (Ovol1 and Ovol2 knockout), and Ovol2 knockout.

$$\frac{[K1]}{dt} =(1 - f_{K14})\mu_{K14}[K14] + (2f_{K1} - 1)\mu_{K1}[K1] \tag{4.19}$$

$$\frac{[Lor]}{dt} =(1 - f_{K1})\mu_{K1}[K1] - d_{Lor}[Lor] \tag{4.20}$$

4.2.3 Estimate the parameters in the mathematical model driven by experimental data

The above mathematical model is built based on the gene regulatory network and the cell lineage. However, the model is not determined before the parameters are set. The following observations be used for determining the parameters in the model.

Both fly Ovo and mammalian Ovol1 reside downstream of key developmental signaling pathways such as Wg/Wnt and BMP/TGFβ [81, 28, 103, 113]. The previous study showed that the Ovol (particularly Ovol1 and Ovol2 based on their expression in skin) involvement in epidermal development using single knockout approaches. Germline ablation of Ovol1 results in a thickened epidermis at birth with expanded spinous layers [24, 103]. The intermediate cells in Ovol1-deficient embryos fail to undergo proliferation arrest, and Ovol1-deficient keratinocytes do not respond to TGFβ signaling and exit cell cycle [103]. Germline ablation of Ovol2 results in mid-gestation lethality, precluding the analysis of epidermal development which occurs afterwards. Analysis of early mutant embryos revealed an over-specification of neural fate at the cost of surface ectodermal fate [91], suggesting a role for Ovol2 in the very early stages of surface epithelial development. The previous study shows that Ovol2 depletion leads to a transient cell expansion but a loss of cells with long-term proliferation potential. In summary, we summarized the experimental data from the previous literatures of the thickness of the three layers of epidermis from E15.5 to E18.5 at the different conditions, Wildtype(WT), Ovol1 deficient (Ovol-/-), Ovol2 overexpressed (Ovol2BT), Ovol1 and Ovol2 knockout (DKO), and Ovol2 knockout (Ovol2SSKO), in Figure 4.2. The blank ones are the missing data, which couldn't be found in the literatures.

We estimate the parameters in the model by the following objective function.

$$J(\Theta) = \sum_{i=1}^{5} \sum_{j=1}^{4} \omega_i ||V^{(\mathrm{th})}(\Theta; t_j^{(i)}) - V^{(\mathrm{exp})}(t_j^{(i)})||^2 \tag{4.21}$$

In the above objective function, $|| \bullet ||$ denotes the L^2 norm operator; Θ denotes the vector of parameters in the lineage model;In the first summation, there are 5 conditions, and in the second summation, there are 4 time points. ω_i represents the corresponding coefficients of the weights in the 5 conditions and here we set all of them as 1 equally based on the

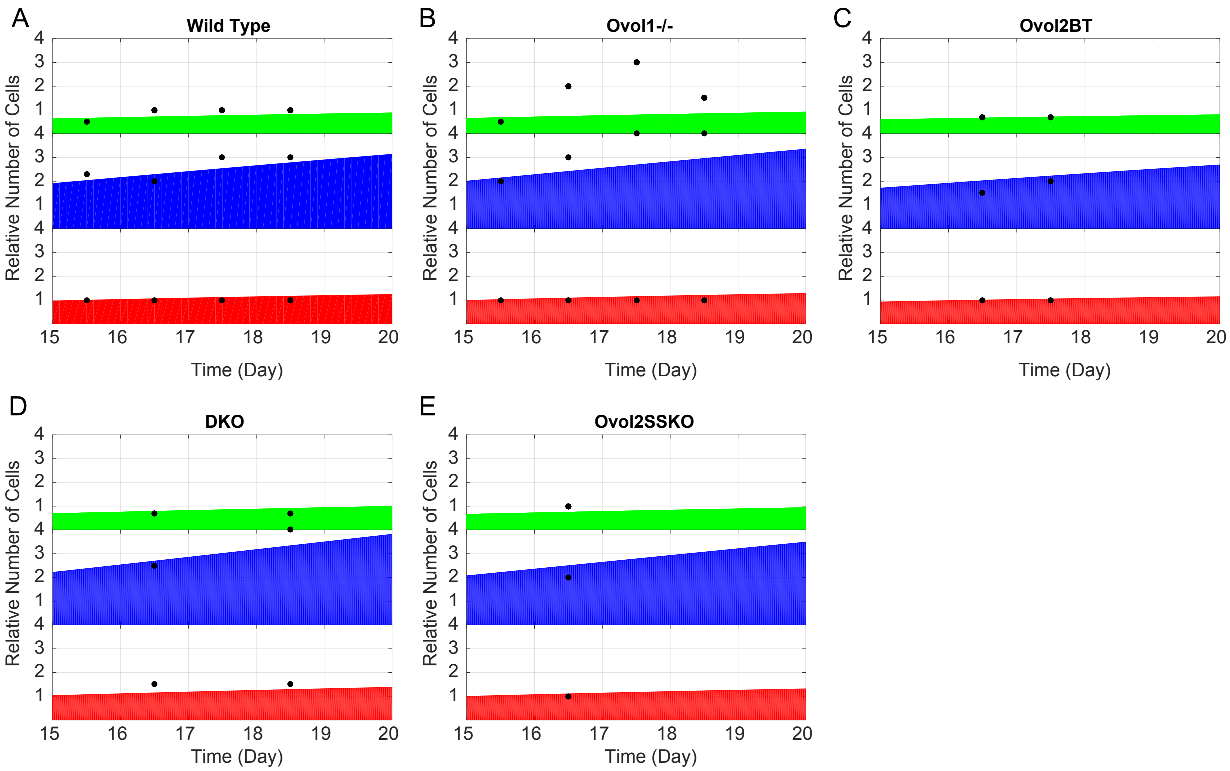


Figure 4.3: The simulation results of skin epidermis thickness with three layers at different stages under different conditions.The green area is the thickness of the granular layer. The blue area is the thickness of the spinous layer. The red area is the thickness of the basal layer.The black dot is the experimental data.

equal importance of the experimental conditions. In addition, V^{th} and exp represent the theoretical and experimental vectors of the observations composed of three types of cells. Finally, a total of 39 system outputs (experimental data) are therefore used to fit 42 model parameters.

All fitting results are shown in Figure 4.3. Most of the fitting errors are rather small except for those under the condition of Ovol1 knockout. There are many reasons resulting in the greater errors. One is that the experimental noise.

4.2.4 Parameter sensitivity analysis of the fitting model

Parameter sensitivity analysis is used to determine the effect quantitatively that specific parameters have on the outputs. The sensitivity coefficient of the parameter P is defined as:

$$S_P^i = \frac{\partial L_i/L_i}{\partial P/P} \cong \frac{\triangle L_i/L_i}{\triangle P/P} \tag{4.22}$$

Where L is the system output including the relative number of cells in the three layers, basal layer, spinous layer, and the granular layer. P is the one of the parameters in the fitting model. Individual parameters were perturbed by 1% from their estimated values resulting in the changes in the system output $\triangle L_i$. Essentially the sensitivity coefficient denotes the percentage change of output caused by perturbing a parameter P. All of the sensitivity coefficients are shown in Figure 4.4. We can see that regulations from TGFβ to Zeb1, from Notch to p63, and from Ovol2 to p63 have high sensitivity coefficients. In the following, we will discuss the functions of Ovol family, Zeb1, and p63 one by one.

4.2.5 Ovol1 and Ovol2 inhibit the development of mouse embryonic epidermis

The simulation results show that Ovol1 and Ovol2 inhibit mouse embryonic epidermis development. Figure 4.5A to Figure 4.5C are the relative number of cells in the three layers when Ovol1 or Ovol2 are overexpressed. Then we can see that the inhibition of development by Ovol2 overexpressed is more severe than the one by overexpressed Ovol1 and the inhibition is a synergy effect in Figure 4.5D. Similarly, the same conclusion could obtain from Ovol1 or Ovol2 knockout from Figure 4.5E to Figure 4.5G.

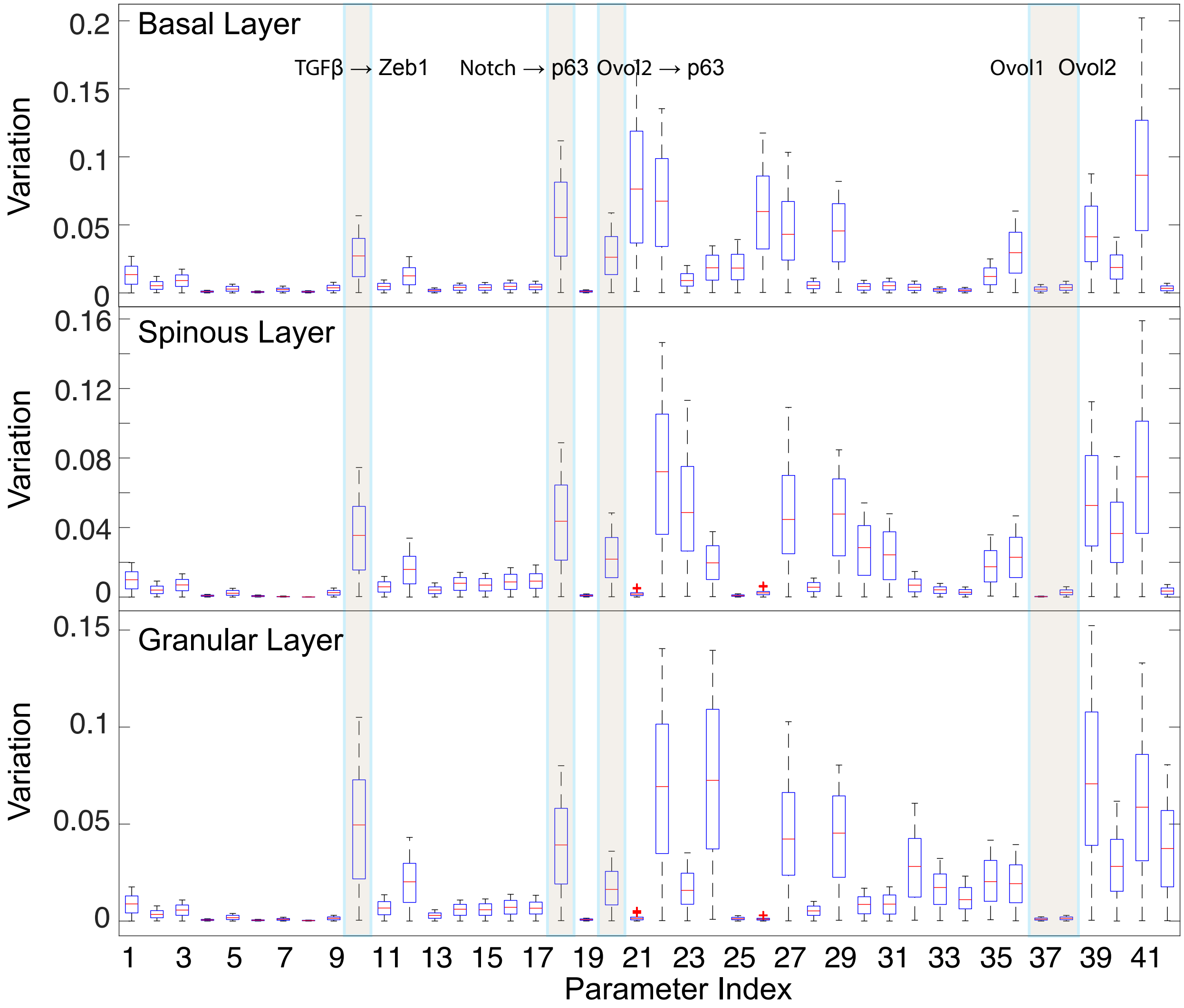


Figure 4.4: The sensitivity analysis results for each parameter in the mathematical model. The x-axis is the parameter index.

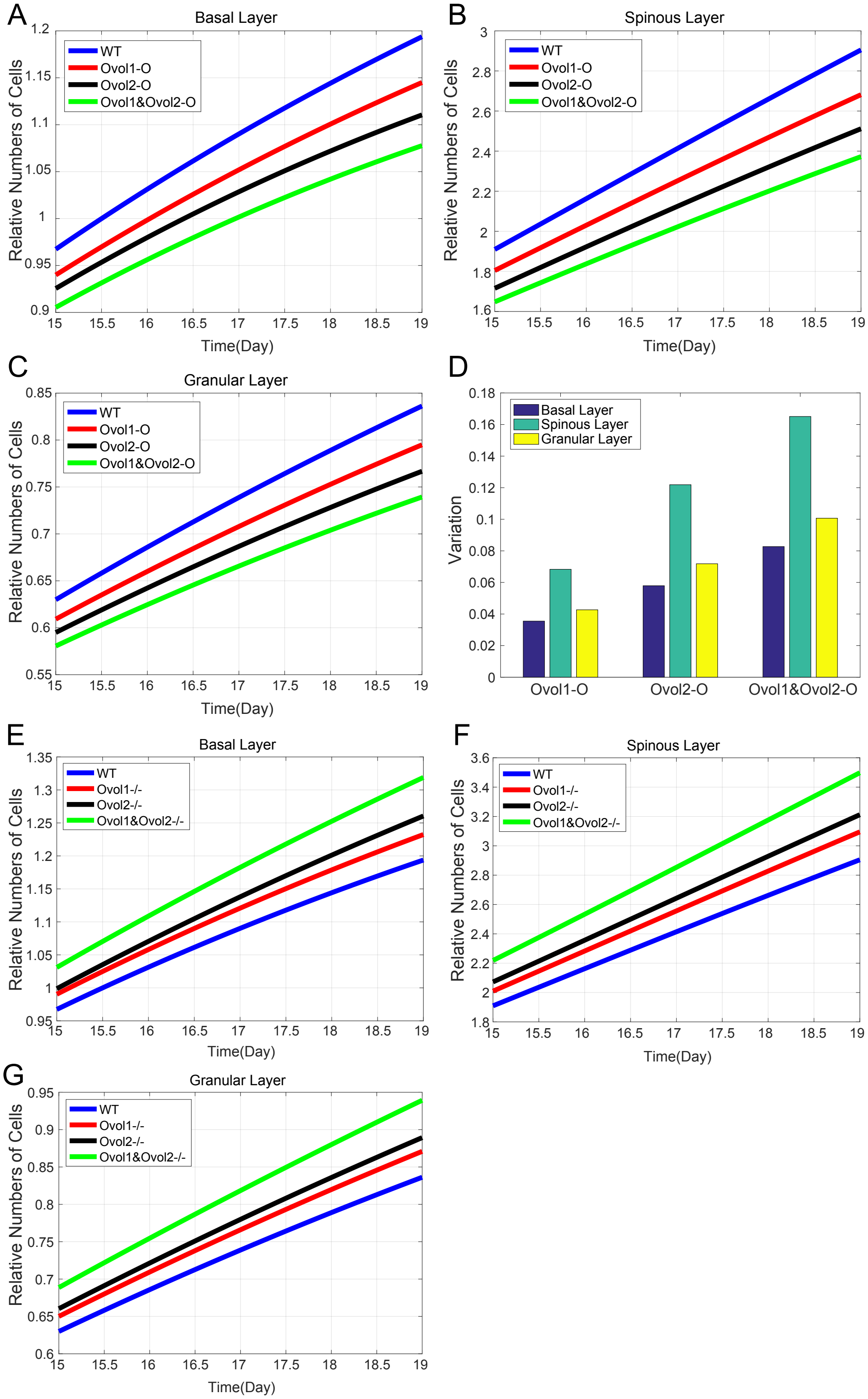


Figure 4.5: The embryonic epidermal development under the mutation of Ovol1 or Ovol2.

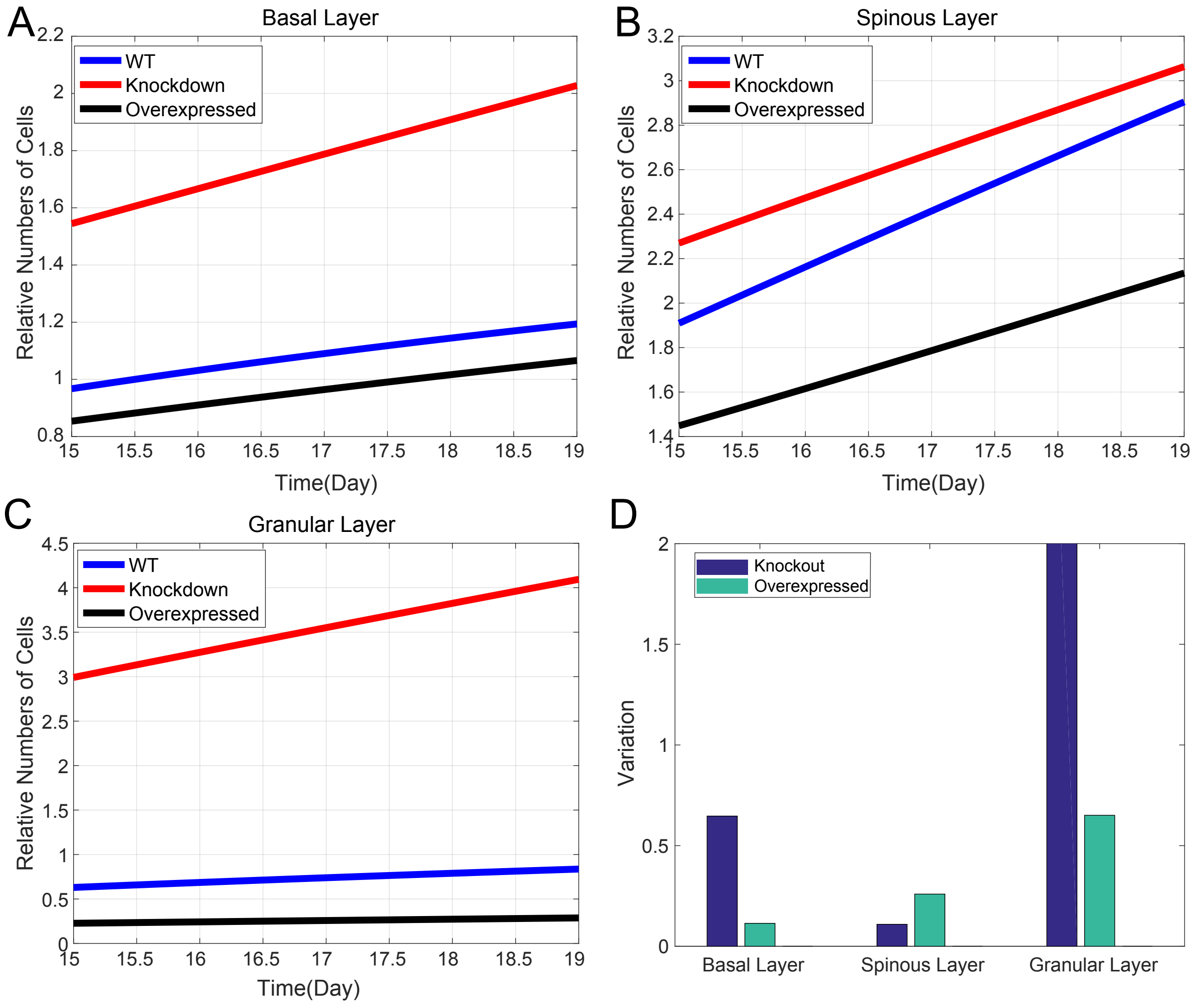


Figure 4.6: The embryonic epidermal development under the mutation of Zeb1.

4.2.6 Zeb1 inhibit the development of mouse embryonic epidermis

The simulation results show that Zeb1 inhibit mouse embryonic epidermis development. Figure 4.6A to Figure 4.6C are the relative number of cells in the three layers when Zeb1 is overexpressed or knockout.

4.2.7 p63 promotes the development of mouse embryonic epidermis

The simulation results show that p63 promotes mouse embryonic epidermis development. Figure 4.7A to Figure 4.7C are the relative number of cells in the three layers when p63 is overexpressed or knockout. Especially p63 is knockout, then mouse embryonic epidermis cannot develop. It is lethal for mouse embryonic epidermis development. It is consistent with the previous literatures [73, 98].

4.3 Conclusion

We constructed a multi-scale mathematical model for mouse embryonic epidermis development integrating the gene network and the cell lineage. The regulatory relationship between gene was based on the previous literatures and the three stages cell lineage represented the three layers of the epidermis. The parameters are estimated to use the simulation outputs to fit the experimental data. The model predicts that Ovol1 and Ovol2 inhibit the epidermis development. p63 promotes the development of embryonic epidermis development which is the consistent with the results in the previous literatures.

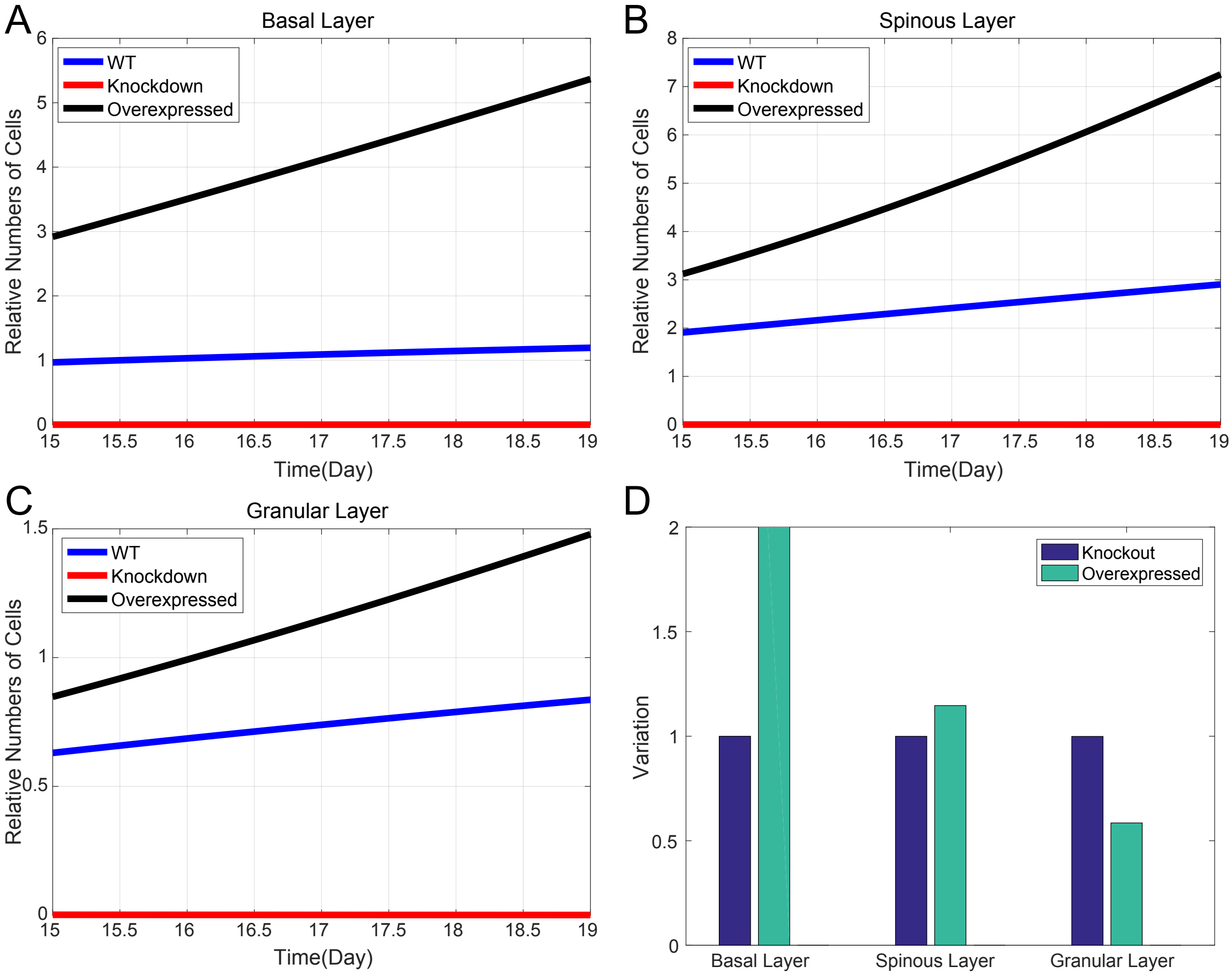


Figure 4.7: The embryonic epidermal development under the mutation of p63.

Chapter 5

Network inference integrating prior information

[Chapeter 5 is an ongoing project. Tao Peng, Qing Nie. Network inference incorporating the prior information.]

5.1 Introduction

More and more methodologies have been developed to infer gene regulatory networks from gene expression data such as graphical models, information-theoretic approaches, and ordinary differential equations. However, it is a great challenge to integrate the information from other measurements to infer the gene regulatory network. These technologies include incorporating DNA motif sequence in gene promoter regions [118, 131, 143], combining multiple microarray datasets from the same organism across multiple experiments [100, 157] or from completely different organisms [36], and integrating proteomics and metabolomics [144]. ChIP-chip and ChIP-seq are used to detect the physical interactions between different genes and they have been employed to construct putative regulatory networks. However, there is no research integrating this type of data to the inference of network by genome-wide expression data [85]. Since the data is limited by the stable antibodies, it is impossible to construct the interactions between different genes. It leads that we have to use the data of physical interactions as the prior information to infer the network of genes. The previous work has shown that various types of experimental data can be formulated into the framework utilizing regularization parameters to take the prior information into account [22]. In the following study, we will study the model selections when we have different kinds of prior information.

5.2 Mathematical model

We model the gene network inference as the following linear ordinary differential equation (ODEs).

$$\frac{x_i(t)}{dt} = G_i(Y(t)) - r_i x_i(t) \tag{5.1}$$

where $x_i(t) \quad (i = 1, 2, ..., n)$ is the expression level of target gene i at time t. n is the number of target genes in the gene network. r_i is the regulatory coefficient of gene i itself. $Y(t)$ is the vector $(y_1, y_2, ..., y_m)^T$. $y_j(t) \quad (j = 1, 2, ..., n)$ is the expression level of regulatory gene j at time t. $\frac{dx_i(t)}{dt} \quad (i = 1, 2, .., n)$ is the expression rate of gene i. G_i is the effect of all regulatory genes on the gene i expression rate. The effect includes transcription regulation, translation regulation, post-translation modification and so on. For current simplicity of presentation we will take the form that approximate the gene regulatory network with a linear system of equations.

$$G_i(Y(t)) \approx \sum_{j=1}^{m} M_{ij} y_j(t) \tag{5.2}$$

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{m} M_{ij} y_j(t) - r_i x_i(t) \tag{5.3}$$

Then we can rewrite the ODEs.

$$\frac{dX(t)}{dt} = MY(t) - X(t)R \tag{5.4}$$

where $X(t) = (x_1(t), x_2(t), ..., x_n(t))^T$, $R = (r_1, r_2, ..., r_n)$, regulatory matrix $M = (M_{ij})$. Solving Eqs 5.4 can be expressed as the solution of least-squares minimization problem given N observations, $(Y^1, X^1), (Y^2, X^2), ..., (Y^N, X^N)$ under the steady states.

$$M^* = arg \min_M f(M) = arg \min_M \sum_{j=1}^{m} ||MY^j - X^j R|| \tag{5.5}$$

Then we can rewrite Equation (5.5) as the classical least-square problem.

$$M^* = arg \min_W f(W) = arg \min_W \sum_{j=1}^{m} ||WY^j - X^j|| \tag{5.6}$$

5.2.1 Enforcing sparse regulatory matrix

$$M^* = arg \min_W g(W) = arg \min_W \sum_{j=1}^{m} ||WY^j - X^j|| + \alpha ||W||_1 \tag{5.7}$$

where $W = (w_{ij})$ and $||W||_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} |w_{ij}|$. α is learned through cross-validation with larger values for α producing a more sparse matrix while $\alpha = 0$ corresponds to the standard least-squares regression problem.

5.2.2 Including prior network information

Existing network information can be incorporated into the minimization problem by adding an additional constraint for connections in the network. Given a W^0 matrix with positive entries $W^0_{ij} \geq 0$ indicating the lack of interaction for regulatory gene j on regulatory gene i, the problem becomes:

$$M^* = arg \min_W g(W) = arg \min_W \sum_{j=1}^{m} ||WY^j - X^j|| + \alpha||W||_1 + \beta||W \circ W^0|| \qquad (5.8)$$

5.3 Results

5.3.1 Analysis of the prior information of transcription factor regulation information

The gene regulation links are sparse in the gene network. Figure 5.1A shows that the correlation matrix of 104 genes. From this figure we can see that 80% dots are around 0. It means most correlation values are very small.

The number of overlapping links from the inferred gene network and transcription factor regulation network is rather small. Figure 5.1B illustrates that there are only 78 overlapping links when the inferred gene network consists of 3500 links. The overlapping links are 30% of 289 links in the transcription factor regulation network. In order to illustrate the significance of overlapping links we generate 1000 random matrices with different numbers of links shown in Figure 5.1C and find that the numbers of overlapping links of the random matrices are the same with the ones of inferred gene network. The inverse covariance matrix can illustrate the dependence of different genes (Figure 5.1D). It shows that the overlapping links are around 90 when there are 3500 links in the inferred gene networks. It is consistent with the

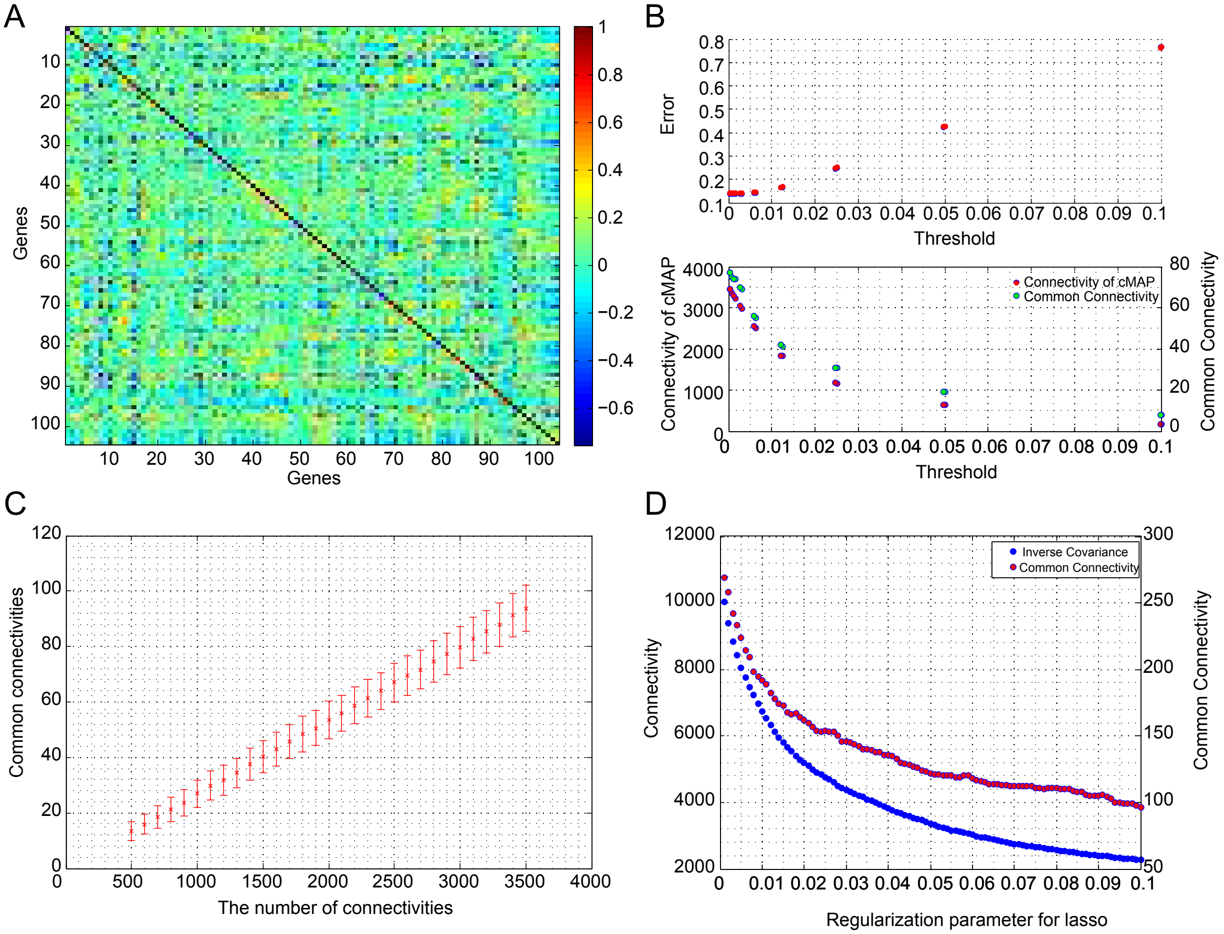


Figure 5.1: Analysis of the significance of prior information. (A) the correlation of genes. (B) upper, the relationship between training error and thresholds, and between test error and thresholds respectively. Lower, the relationship between connectivity and thresholds, and between the overlapping connectivity and thresholds. The overlapping connectivity means that links are shared by the inferred gene network and transcription factor regulation network. (C) the overlapping connectivity between the random gene networks and transcription factor regulation network. (D) the overlapping connectivity from inverse covariance matrix and transcription factor regulation network.

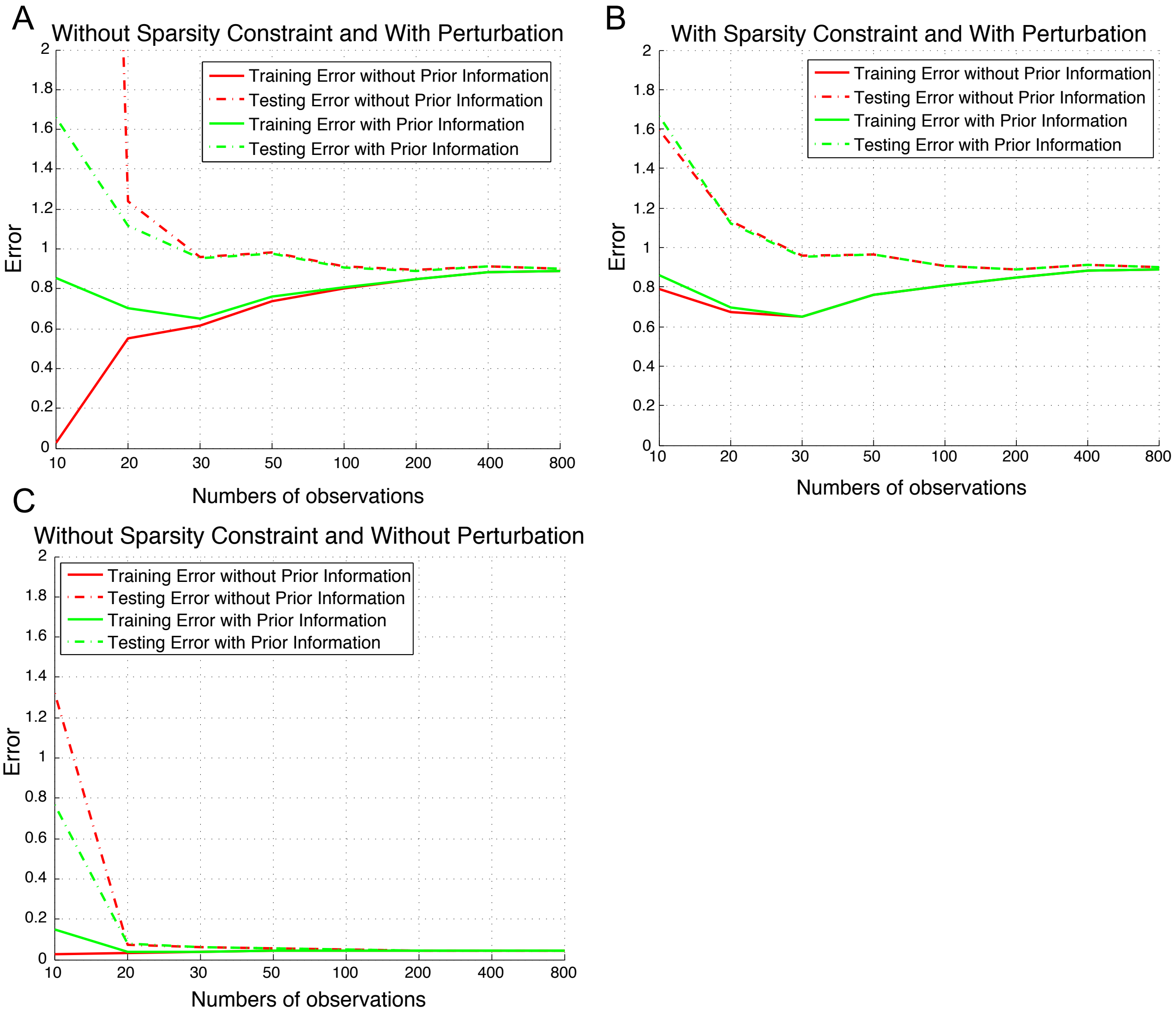


Figure 5.2: The training errors and testing errors using different training methods and different datasets to infer 10 genes network. A. The training errors and testing errors are calculated without sparse constraints. There is perturbation of input data.B. The training errors and testing errors are calculated with sparse constraints. There is perturbation of input data. C. The training errors and testing errors are calculated without sparse constraints. There is no perturbation of input data.

inferred matrix obtained from the algorithm. All show that the prior information provides no contribution on inferring the gene network by large-scale gene expression data.

5.3.2 Prior network information has no effect on the testing error based on a model with sparsity constraint

In the context no sparsity constraint means =0 holds during the optimization. No prior information means =0 is kept during the optimization. The prior information is defined as a matrix, the elements of which are 0 when the elements are mutant. (Here the notation is same with Scotts paper.). In Figure 5.2B, Figure 5.3B, and Figure 5.3D, we can see that there are no changes of testing errors between with prior information and without prior information. It also satisfies the training errors.

5.3.3 Prior network information improves the testing error for a model without sparsity constraint when the number of observations is relatively small

Figure 5.2A, Figure 5.3A, and Figure 5.3C can support these results. In Figure 5.2A the testing errors for the model with prior information are smaller than the ones without prior information when the number of observation is less than 30. The testing errors and training errors tend to approach the same level as the number of observations increases. In Figure 5.3A and Figure 5.3C the testing errors for the model with prior information are smaller than the ones without prior information when the number of observation is less than 200.

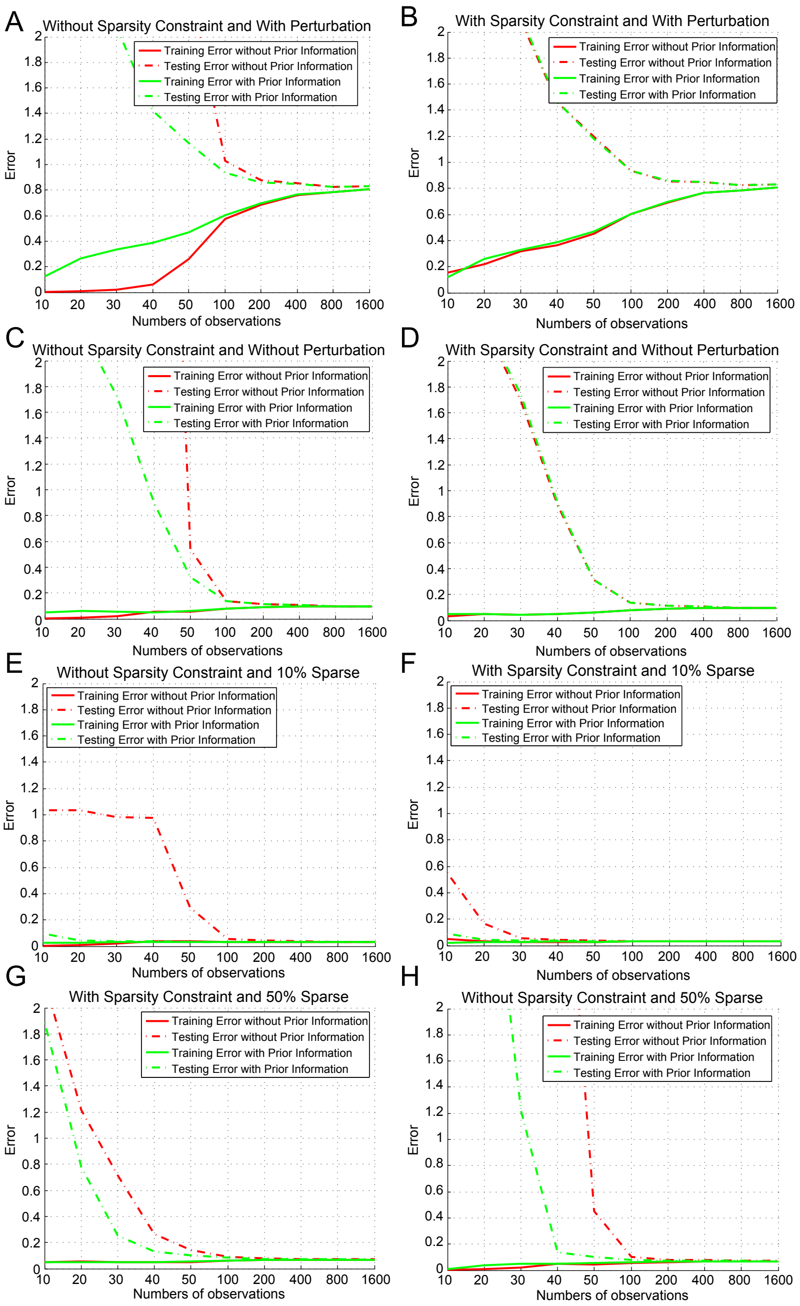


Figure 5.3: The training errors and testing errors using different training methods and different datasets to infer simulated 40 genes network. A-B. The training errors and testing errors are calculated with sparse constraints and without sparse constraints using the perturbed training data from dense transit matrix respectively. C-D. The training errors and testing errors are calculated with sparse constraints and without sparse constraints using the unperturbed training data from dense transit matrix respectively. E-F. The training errors and testing errors are calculated with sparse constraints and without sparse constraints using the unperturbed training data from 10% sparse transit matrix respectively. G-H. The training errors and testing errors are calculated with sparse constraints and without sparse constraints using the unperturbed training data from 50% sparse transit matrix respectively.

5.3.4 Sparse constraints have no effect on the testing error based on a model with prior network information

Figure 5.3E to Figure 5.3H illustrate the results. In Figure 5.3E and Figure 5.3F there are no changes of the testing errors of the model with sparse constraints or without sparse constraints when the prior network information is involved in the model.

5.3.5 Prior network information always improves the test errors with sparse constraints or without sparse constraints

Figure 5.3E to Figure 5.3H can show the results. The prior network information can facilitate the testing errors for the model with sparse constraints. In addition, the network connect matrix W^* is sparser, the prior information is more important for the network inference.

5.3.6 Prior network information has no effect on the testing error based on a model with sparsity constraint

In Figure 5.4, 10 genes are selected randomly and ChIP-seq data of the genes can be from ENCODE database. The following figures show that the prior information can improve the testing error performance when the size of observations is small, such as 100 for 10 genes network inference. However, the prior information cannot benefit the testing error when sparse constraints are involved in the model. All models converge at the same level finally. In Figure 5.5 we randomly employ 20 genes or 40 genes and obtain the same conclusion above.

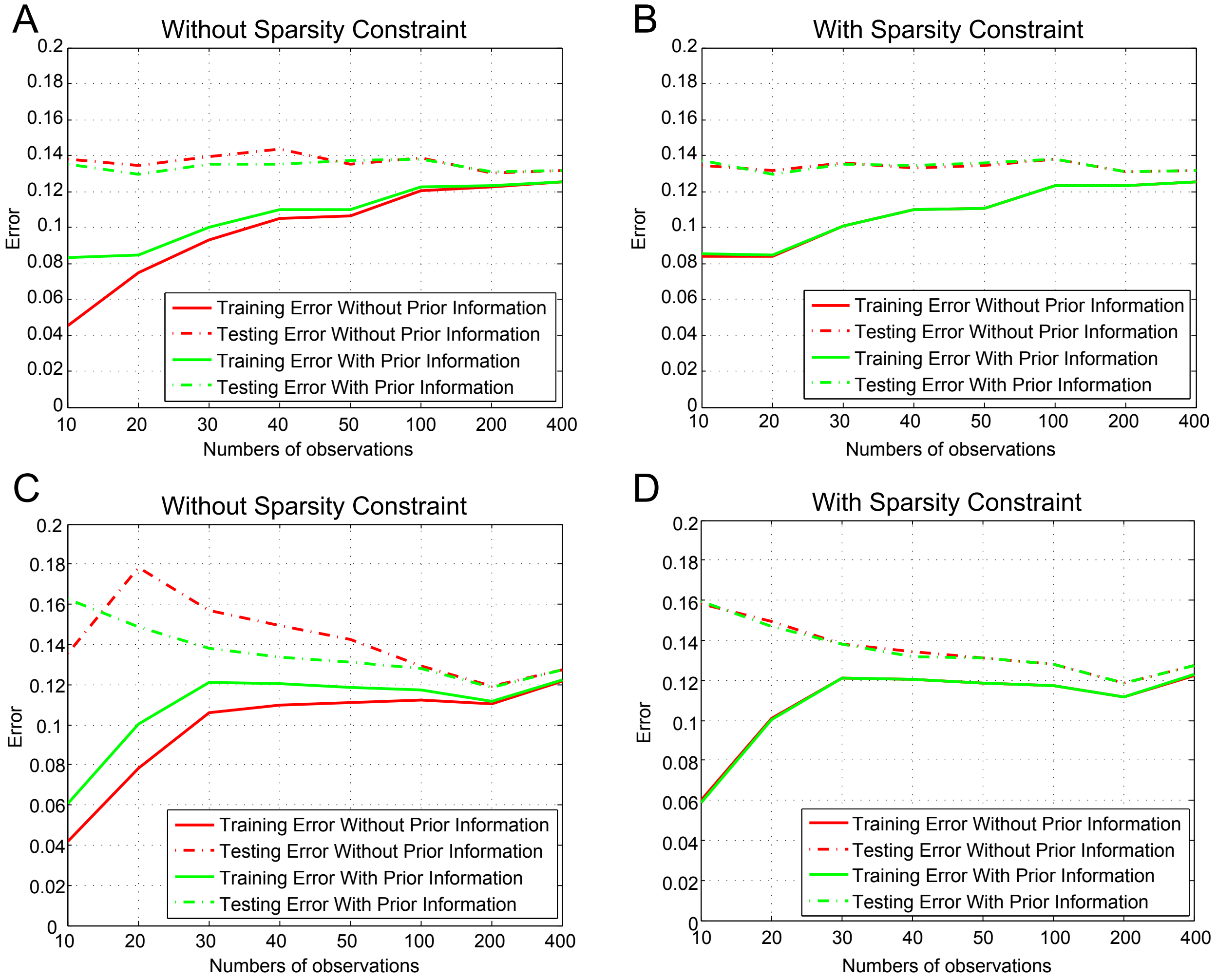


Figure 5.4: Gene network inference using cMAP microarray data and ChIP-seq data. A-B. The training errors and testing errors are calculated with sparse constraints and without sparse constraints based on 10 genes randomly selected from ChIP-seq data. C-D. The training errors and testing errors are calculated with sparse constraints and without sparse constraints based on other 10 genes randomly selected from ChIP-seq data.

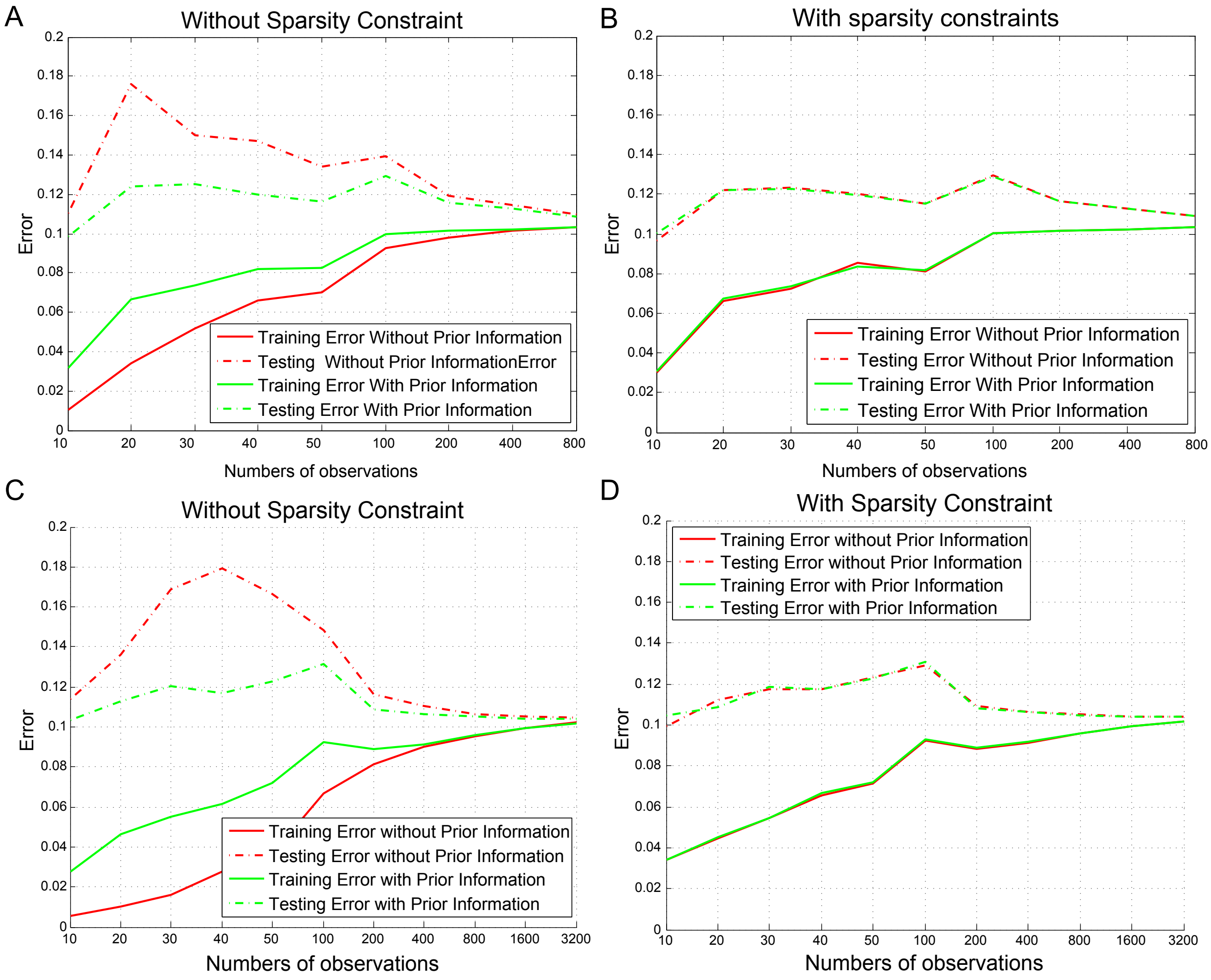


Figure 5.5: Gene network inference using cMAP microarray data and ChIP-seq data. A-B. The training errors and testing errors are calculated with sparse constraints and without sparse constraints based on 20 genes randomly selected from ChIP-seq data. C-D. The training errors and testing errors are calculated with sparse constraints and without sparse constraints based on other 40 genes randomly selected from ChIP-seq data.

5.4 Conclusion

In this chapter we discuss that how the prior information determine the selections of the models. We concludes that the prior information of ChIP-seq data from ENCODE has no effect on the network inference. It might provide a way to check if the prior information is consistent with the information embedding in the expression data. It is an ongoing project and we will continue to finish it in the following.

Bibliography

[1] M. Adler, A. Mayo, and U. Alon. Logarithmic and power law input-output relations in sensory systems with fold-change detection. *PLoS Comput Biol*, 10(8):e1003781, 2014.

[2] R. J. Akhurst, F. Fee, and A. Balmaint. Localized production of tgf-β mrna in tumour promoter-stimulated mouse epidermis. 1988.

[3] C. Alarcón, A.-I. Zaromytidou, Q. Xi, S. Gao, J. Yu, S. Fujisawa, A. Barlas, A. N. Miller, K. Manova-Todorova, M. J. Macias, et al. Nuclear cdks drive smad transcriptional activation and turnover in bmp and tgf-β pathways. *Cell*, 139(4):757–769, 2009.

[4] E. Allen. *Modeling with Itô stochastic differential equations*, volume 22. Springer Science & Business Media, 2007.

[5] E. L. Allgower and K. Georg. *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media, 2012.

[6] E.-a. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er. visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 31(6):545–552, 2013.

[7] A. J. Apicella, I. R. Wickersham, H. S. Seung, and G. M. Shepherd. Laminarly orthogonal excitation of fast-spiking and low-threshold-spiking interneurons in mouse motor cortex. *Journal of Neuroscience*, 32(20):7021–7033, 2012.

[8] L. Azzolin, T. Panciera, S. Soligo, E. Enzo, S. Bicciato, S. Dupont, S. Bresolin, C. Frasson, G. Basso, V. Guzzardo, et al. Yap/taz incorporation in the β-catenin destruction complex orchestrates the wnt response. *Cell*, 158(1):157–170, 2014.

[9] W. L. Baker. A review of models of landscape change. *Landscape ecology*, 2(2):111–133, 1989.

[10] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Peer. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.

[11] B. P. Bernabé, S. Shin, P. D. Rios, L. J. Broadbelt, L. D. Shea, and S. K. Seidlits. Dynamic transcription factor activity networks in response to independently altered mechanical and adhesive microenvironmental cues. *Integrative Biology*, 8(8):844–860, 2016.

[12] C. Blanpain and E. Fuchs. p63: revving up epithelial stem-cell potential. *Nature cell biology*, 9(7):731–733, 2007.

[13] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160, 2015.

[14] F. Buettner and F. J. Theis. A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, 28(18):i626–i632, 2012.

[15] D. P. Burke and D. J. Kelly. Substrate stiffness and oxygen as regulators of stem cell differentiation during skeletal tissue regeneration: a mechanobiological model. *PLoS One*, 7(7):e40737, 2012.

[16] E. M. Callaway. Local circuits in primary visual cortex of the macaque monkey. *Annual review of neuroscience*, 21(1):47–74, 1998.

[17] G. Chalancon, C. N. Ravarani, S. Balaji, A. Martinez-Arias, L. Aravind, R. Jothi, and M. M. Babu. Interplay between gene expression noise and regulatory network architecture. *Trends in genetics*, 28(5):221–232, 2012.

[18] C. Chen, K. Zhang, H. Feng, M. Sasai, and J. Wang. Multiple coupled landscapes and non-adiabatic dynamics with applications to self-activating genes. *Physical Chemistry Chemical Physics*, 17(43):29036–29044, 2015.

[19] K.-C. Chen, T.-Y. Wang, H.-H. Tseng, C.-Y. F. Huang, and C.-Y. Kao. A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics*, 21(12):2883–2890, 2005.

[20] M. Chen, L. Wang, C. C. Liu, and Q. Nie. Noise attenuation in the on and off states of biological switches. *ACS synthetic biology*, 2(10):587–593, 2013.

[21] B. Choi, K.-S. Park, J.-H. Kim, K.-W. Ko, J.-S. Kim, D. K. Han, and S.-H. Lee. Stiffness of hydrogels regulates cellular reprogramming efficiency through mesenchymal-to-epithelial transition and stemness markers. *Macromolecular bioscience*, 16(2):199–206, 2016.

[22] S. Christley, Q. Nie, and X. Xie. Incorporating existing network information into gene network inference. *PloS one*, 4(8):e6799, 2009.

[23] T. R. Cox and J. T. Erler. Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. *Disease models & mechanisms*, 4(2):165–178, 2011.

[24] X. Dai, C. Schonbaum, L. Degenstein, W. Bai, A. Mahowald, and E. Fuchs. The ovo gene required for cuticle formation and oogenesis in flies is involved in hair formation and spermatogenesis in mice. *Genes & development*, 12(21):3452–3463, 1998.

[25] J. Dantzker and E. Callaway. Laminar sources of synaptic input to cortical inhibitory interneurons and pyramidal neurons. *Nature neuroscience*, 3(7):701–707, 2000.

[26] S. B.-T. de Leon and E. H. Davidson. Modeling the dynamics of transcriptional gene regulatory networks for animal development. *Developmental biology*, 325(2):317–328, 2009.

[27] L. de Vargas Roditi and M. Claassen. Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. *Current opinion in biotechnology*, 34:9–15, 2015.

[28] P. Descargues, A. K. Sil, Y. Sano, O. Korchynskyi, G. Han, P. Owens, X.-J. Wang, and M. Karin. Ikkα is a critical coregulator of a smad4-independent tgfβ-smad2/3 signaling pathway that controls keratinocyte differentiation. *Proceedings of the National Academy of Sciences*, 105(7):2487–2492, 2008.

[29] A. Dhooge, W. Govaerts, Y. A. Kuznetsov, H. Meijer, and B. Sautois. New features of the software matcont for bifurcation analysis of dynamical systems. *Mathematical and Computer Modelling of Dynamical Systems*, 14(2):147–175, 2008.

[30] P. D. P. Dingal, A. M. Bradshaw, S. Cho, M. Raab, A. Buxboim, J. Swift, and D. E. Discher. Fractal heterogeneity in minimal matrix models of scars modulates stiff-niche stem-cell responses via nuclear exit of a mechanorepressor. *Nature materials*, 14(9):951–960, 2015.

[31] R. J. Douglas and K. A. Martin. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–451, 2004.

[32] S. Dupont, L. Morsut, M. Aragona, E. Enzo, S. Giulitti, M. Cordenonsi, F. Zanconato, J. Le Digabel, M. Forcato, S. Bicciato, et al. Role of yap/taz in mechanotransduction. *Nature*, 474(7350):179–183, 2011.

[33] J. A. Damour and R. C. Froemke. Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron*, 86(2):514–528, 2015.

[34] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[35] A. J. Engler, S. Sen, H. L. Sweeney, and D. E. Discher. Matrix elasticity directs stem cell lineage specification. *Cell*, 126(4):677–689, 2006.

[36] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biol*, 5(1):e8, 2007.

[37] E. Fino and R. Yuste. Dense inhibitory connectivity in neocortex. *Neuron*, 69(6):1188–1203, 2011.

[38] D. V. Foster, J. G. Foster, S. Huang, and S. A. Kauffman. A model of sequential branching in hierarchical cell fate determination. *Journal of theoretical biology*, 260(4):589–597, 2009.

[39] T. Frum and A. Ralston. Cell signaling and transcription factors regulating cell fate during formation of the mouse blastocyst. *Trends in Genetics*, 31(7):402–410, 2015.

[40] A. Gandarillas and F. M. Watt. c-myc promotes differentiation of human epidermal stem cells. *Genes & development*, 11(21):2869–2882, 1997.

[41] F. Gattazzo, A. Urciuolo, and P. Bonaldo. Extracellular matrix: a dynamic microenvironment for stem cell niche. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1840(8):2506–2519, 2014.

[42] R. Gesztelyi, J. Zsuga, A. Kemeny-Beke, B. Varga, B. Juhasz, and A. Tosaki. The hill equation and the origin of quantitative pharmacology. *Archive for history of exact sciences*, 66(4):427–438, 2012.

[43] P. M. Gilbert, K. L. Havenstrite, K. E. Magnusson, A. Sacco, N. A. Leonardi, P. Kraft, N. K. Nguyen, S. Thrun, M. P. Lutolf, and H. M. Blau. Substrate elasticity regulates skeletal muscle stem cell self-renewal in culture. *Science*, 329(5995):1078–1081, 2010.

[44] A. D. Goldberg, C. D. Allis, and E. Bernstein. Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638, 2007.

[45] F. Guilak, D. M. Cohen, B. T. Estes, J. M. Gimble, W. Liedtke, and C. S. Chen. Control of stem cell fate by physical interactions with the extracellular matrix. *Cell stem cell*, 5(1):17–26, 2009.

[46] G. Guo, M. Huss, G. Q. Tong, C. Wang, L. L. Sun, N. D. Clarke, and P. Robson. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell*, 18(4):675–685, 2010.

[47] L. Haghverdi, F. Buettner, and F. J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, page btv325, 2015.

[48] B. Haider, M. Häusser, and M. Carandini. Inhibition dominates sensory responses in the awake cortex. *Nature*, 493(7430):97–100, 2013.

[49] G. Halder, S. Dupont, and S. Piccolo. Transduction of mechanical and cytoskeletal cues by yap and taz. *Nature reviews Molecular cell biology*, 13(9):591–600, 2012.

[50] S.-i. Harada and G. A. Rodan. Control of osteoblast function and regulation of bone mass. *Nature*, 423(6937):349–355, 2003.

[51] K. D. Harris and G. M. Shepherd. The neocortical circuit: themes and variations. *Nature neuroscience*, 18(2):170–181, 2015.

[52] A. Hasenstaub, Y. Shu, B. Haider, U. Kraushaar, A. Duque, and D. A. McCormick. Inhibitory postsynaptic potentials carry synchronized frequency information in active cortical networks. *Neuron*, 47(3):423–435, 2005.

[53] G. H. Heppner. Tumor heterogeneity. *Cancer research*, 44(6):2259–2265, 1984.

[54] D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

[55] W. R. Holmes and Q. Nie. Interactions and tradeoffs between cell recruitment, proliferation, and differentiation affect cns regeneration. *Biophysical journal*, 106(7):1528–1536, 2014.

[56] J.-H. Hong, E. S. Hwang, M. T. McManus, A. Amsterdam, Y. Tian, R. Kalmukova, E. Mueller, T. Benjamin, B. M. Spiegelman, P. A. Sharp, et al. Taz, a transcriptional modulator of mesenchymal stem cell differentiation. *Science*, 309(5737):1074–1078, 2005.

[57] J.-H. Hong and M. B. Yaffe. Taz: a β-catenin-like molecule that regulates mesenchymal stem cell differentiation. *Cell cycle*, 5(2):176–179, 2006.

[58] G. Hu. Stochastic forces and nonlinear systems. *Shanghai Scientific and Technological Education Publishing House, Shanghai*, page 184, 1994.

[59] T. Ikrar, N. D. Olivas, Y. Shi, and X. Xu. Mapping inhibitory neuronal circuits by laser scanning photostimulation. *JoVE (Journal of Visualized Experiments)*, (56):e3109–e3109, 2011.

[60] T. Ikrar, Y. Shi, T. Velasquez, M. Goulding, and X. Xu. Cell-type specific regulation of cortical excitability through the allatostatin receptor system. *Frontiers in neural circuits*, 6:2, 2012.

[61] B. Ingalls. Mathematical modelling in systems biology: An introduction. *Internet.[cited at p. 117]*, 2013.

[62] I. L. Ivanovska, J.-W. Shin, J. Swift, and D. E. Discher. Stem cell mechanobiology: diverse lessons from bone marrow. *Trends in cell biology*, 25(9):523–532, 2015.

[63] H. Jeong, S. Bae, S. Y. An, M. R. Byun, J.-H. Hwang, M. B. Yaffe, J.-H. Hong, and E. S. Hwang. Taz as a novel enhancer of myod-mediated myogenic differentiation. *The FASEB Journal*, 24(9):3310–3320, 2010.

[64] Z. Ji and H. Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic Acids Research*, page gkw430, 2016.

[65] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: a survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[66] K.-T. Kang, J.-H. Park, H.-J. Kim, H.-M. Lee, K.-I. Lee, H.-H. Jung, H.-Y. Lee, S. Young-Bock, and J. W. Jang. Study of tissue differentiation of mesenchymal stem cells by mechanical stimuli and an algorithm for bone fracture healing. *Tissue Engineering and Regenerative Medicine*, 8(4):359–370, 2011.

[67] C. Kapfer, L. L. Glickfeld, B. V. Atallah, and M. Scanziani. Supralinear increase of recurrent inhibition during sparse activity in the somatosensory cortex. *Nature neuroscience*, 10(6):743–753, 2007.

[68] D. Kätzel, B. V. Zemelman, C. Buetfering, M. Wölfel, and G. Miesenböck. The columnar and laminar organization of inhibitory connections to neocortical excitatory cells. *Nature neuroscience*, 14(1):100–107, 2011.

[69] S. Khetan, M. Guvendiren, W. R. Legant, D. M. Cohen, C. S. Chen, and J. A. Burdick. Degradation-mediated cellular traction directs stem cell fate in covalently crosslinked three-dimensional hydrogels. *Nature materials*, 12(5):458–465, 2013.

[70] K. A. Kilian, B. Bugarija, B. T. Lahn, and M. Mrksich. Geometric cues for directing the differentiation of mesenchymal stem cells. *Proceedings of the National Academy of Sciences*, 107(11):4872–4877, 2010.

[71] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1):1–6, 1998.

[72] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[73] M. I. Koster, S. Kim, A. A. Mills, F. J. DeMayo, and D. R. Roop. p63 is the molecular switch for initiation of an epithelial stratification program. *Genes & development*, 18(2):126–131, 2004.

[74] M. I. Koster and D. R. Roop. Mechanisms regulating epithelial stratification. *Annu. Rev. Cell Dev. Biol.*, 23:93–113, 2007.

[75] S. J. Kuhlman, N. D. Olivas, E. Tring, T. Ikrar, X. Xu, and J. T. Trachtenberg. A disinhibitory microcircuit initiates critical-period plasticity in the visual cortex. *Nature*, 501(7468):543–546, 2013.

[76] D. A. Lawson, N. R. Bhakta, K. Kessenbrock, K. D. Prummel, Y. Yu, K. Takai, A. Zhou, H. Eyob, S. Balakrishnan, C.-Y. Wang, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 2015.

[77] J. A. Lee and M. Verleysen. *Nonlinear dimensionality reduction.* Springer Science & Business Media, 2007.

[78] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 2007.

[79] K. R. Levental, H. Yu, L. Kass, J. N. Lakins, M. Egeblad, J. T. Erler, S. F. Fong, K. Csiszar, A. Giaccia, W. Weninger, et al. Matrix crosslinking forces tumor progression by enhancing integrin signaling. *Cell*, 139(5):891–906, 2009.

[80] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, D. A. El-ad, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[81] B. Li, Q. Dai, L. Li, M. Nair, D. R. Mackay, and X. Dai. Ovol2, a mammalian homolog of drosophila ovo: gene structure, chromosomal mapping, and aberrant expression in blind-sterile mice. *Genomics*, 80(3):319–325, 2002.

[82] C. Li, E. Wang, and J. Wang. Landscape topography determines global stability and robustness of a metabolic network. *ACS synthetic biology*, 1(6):229–239, 2012.

[83] C. Li and J. Wang. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLoS Comput Biol*, 9(8):e1003165, 2013.

[84] C. Li and J. Wang. Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle. *Proceedings of the National Academy of Sciences*, 111(39):14130–14135, 2014.

[85] H. Li and M. Zhan. Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data. *Bioinformatics*, 24(17):1874–1880, 2008.

[86] Q. Li, A. Wennborg, E. Aurell, E. Dekel, J.-Z. Zou, Y. Xu, S. Huang, and I. Ernberg. Dynamics inside the cancer cell attractor reveal cell heterogeneity, limits of stability, and escape. *Proceedings of the National Academy of Sciences*, 113(10):2672–2677, 2016.

[87] S. Liebscher, T. Kirschstein, and C. Becker. The flood algorithma multivariate, self-organizing-map-based, robust location and covariance estimator. *Statistics and Computing*, 22(1):325–336, 2012.

[88] X. Liu, V. Alexander, K. Vijayachandra, E. Bhogte, I. Diamond, and A. Glick. Conditional epidermal expression of tgfβ1 blocks neonatal lethality but causes a reversible hyperplasia and alopecia. *Proceedings of the National Academy of Sciences*, 98(16):9139–9144, 2001.

[89] D. Lü, C. Luo, C. Zhang, Z. Li, and M. Long. Differential regulation of morphology and stemness of mouse embryonic stem cells by substrate stiffness and topography. *Biomaterials*, 35(13):3945–3955, 2014.

[90] H. Lv, L. Li, M. Sun, Y. Zhang, L. Chen, Y. Rong, and Y. Li. Mechanism of regulation of stem cell differentiation by matrix stiffness. *Stem cell research & therapy*, 6(1):103, 2015.

[91] D. R. Mackay, M. Hu, B. Li, C. Rhéaume, and X. Dai. The mouse ovol2 gene is required for cranial neural tube development. *Developmental biology*, 291(1):38–52, 2006.

[92] E. Maharam, M. Yaport, N. L. Villanueva, T. Akinyibi, D. Laudier, Z. He, D. J. Leong, and H. B. Sun. Rho/rock signal transduction pathway is required for msc tenogenic differentiation. *Bone research*, 3:15015, 2015.

[93] E. Marco, R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G.-C. Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 111(52):E5643–E5650, 2014.

[94] H. Markram, M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu. Interneurons of the neocortical inhibitory system. *Nature reviews. Neuroscience*, 5(10):793, 2004.

[95] F. Matsuoka, I. Takeuchi, H. Agata, H. Kagami, H. Shiono, Y. Kiyota, H. Honda, and R. Kato. Morphology-based prediction of osteogenic differentiation potential of human mesenchymal stem cells. *PloS one*, 8(2):e55082, 2013.

[96] R. McBeath, D. M. Pirone, C. M. Nelson, K. Bhadriraju, and C. S. Chen. Cell shape, cytoskeletal tension, and rhoa regulate stem cell lineage commitment. *Developmental cell*, 6(4):483–495, 2004.

[97] M. G. Mendez and P. A. Janmey. Transcription factor regulation by mechanical stress. *The international journal of biochemistry & cell biology*, 44(5):728–732, 2012.

[98] A. A. Mills, B. Zheng, X.-J. Wang, H. Vogel, D. R. Roop, and A. Bradley. p63 is a p53 homologue required for limb and epidermal morphogenesis. *Nature*, 398(6729):708–713, 1999.

[99] V. Moignard, I. C. Macaulay, G. Swiers, F. Buettner, J. Schütte, F. J. Calero-Nieto, S. Kinston, A. Joshi, R. Hannah, F. J. Theis, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature cell biology*, 15(4):363–372, 2013.

[100] F. Mordelet and J.-P. Vert. Sirene: supervised inference of regulatory networks. *Bioinformatics*, 24(16):i76–i82, 2008.

[101] S. J. Mousavi and M. H. Doweidar. Role of mechanical cues in cell differentiation and proliferation: a 3d numerical model. *PloS one*, 10(5):e0124529, 2015.

[102] W. L. Murphy, T. C. McDevitt, and A. J. Engler. Materials as stem cell regulators. *Nature materials*, 13(6):547–557, 2014.

[103] M. Nair, A. Teng, V. Bilanchone, A. Agrawal, B. Li, and X. Dai. Ovol1 regulates the growth arrest of embryonic epidermal progenitor cells and represses c-myc transcription. *J Cell Biol*, 173(2):253–264, 2006.

[104] H. Nguyen, B. J. Merrill, L. Polak, M. Nikolova, M. Rendl, T. M. Shaver, H. A. Pasolli, and E. Fuchs. Tcf3 and tcf4 are essential for long-term homeostasis of skin epithelia. *Nature genetics*, 41(10):1068–1075, 2009.

[105] J. A. Nowak, L. Polak, H. A. Pasolli, and E. Fuchs. Hair follicle stem cells are specified and function in early skin morphogenesis. *Cell stem cell*, 3(1):33–43, 2008.

[106] A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96, 2015.

[107] N. D. Olivas, V. Quintanar-Zilinskas, Z. Nenadic, and X. Xu. Laminar circuit organization and response modulation in mouse visual cortex. *Frontiers in neural circuits*, 6:70, 2012.

[108] S. R. Olsen, D. S. Bortone, H. Adesnik, and M. Scanziani. Gain control by layer six in cortical circuits of vision. *Nature*, 483(7387):47–52, 2012.

[109] S. Openshaw, M. Blake, C. Wymer, et al. Using neurocomputing methods to classify britains residential areas. *Innovations in GIS*, 2:97–111, 1995.

[110] A. M. Packer and R. Yuste. Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? *Journal of Neuroscience*, 31(37):13260–13271, 2011.

[111] M. Partridge, M. Green, J. Langdon, and M. Feldmann. Production of tgf-alpha and tgf-beta by cultured keratinocytes, skin and oral squamous cell carcinomas–potential autocrine regulation of normal and malignant epithelial cell proliferation. *British journal of cancer*, 60(4):542, 1989.

[112] M. J. Paszek, N. Zahir, K. R. Johnson, J. N. Lakins, G. I. Rozenberg, A. Gefen, C. A. Reinhart-King, S. S. Margulies, M. Dembo, D. Boettiger, et al. Tensional homeostasis and the malignant phenotype. *Cancer cell*, 8(3):241–254, 2005.

[113] F. Payre, A. Vincent, and S. Carreno. ovo/svb integrates wingless and der pathways to control epidermis differentiation. *Nature*, 400(6741):271–275, 1999.

[114] T. Peng, L. Liu, A. L. MacLean, C. W. Wong, W. Zhao, and Q. Nie. A mathematical model of mechanotransduction reveals how mechanical memory regulates mesenchymal stem cell fate decisions. *BMC Systems Biology*, 11(1):55, 2017.

[115] T. Peng and Q. Nie. Somsc: Self-organization-map for high-dimensional single-cell data of cellular states and their transitions. *bioRxiv*, page 124693, 2017.

[116] T. Peng, H. Peng, D. S. Choi, J. Su, C.-C. Chang, and X. Zhou. Modeling cell–cell interactions in regulating multiple myeloma initiating cell fate. *IEEE journal of biomedical and health informatics*, 18(2):484–491, 2014.

[117] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[118] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, 29(2):153–159, 2001.

[119] S. Pluta, A. Naka, J. Veit, G. Telian, L. Yao, R. Hakim, D. Taylor, and H. Adesnik. A direct translaminar inhibitory circuit tunes cortical output. *Nature neuroscience*, 2015.

[120] H. Prinz. Hill coefficients, dose–response curves and allosteric mechanisms. *Journal of chemical biology*, 3(1):37–44, 2010.

[121] D. Pyle. *Data preparation for data mining*, volume 1. Morgan Kaufmann, 1999.

[122] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10):886–891, 2011.

[123] V. K. Raghunathan, J. T. Morgan, B. Dreier, C. M. Reilly, S. M. Thomasy, J. A. Wood, I. Ly, B. C. Tuyen, M. Hughbanks, C. J. Murphy, et al. Role of substratum stiffness in modulating genes associated with extracellular matrix and mechanotransducers yap and tazmechanotransduction in trabecular meshwork cells and matrix proteins. *Investigative ophthalmology & visual science*, 54(1):378–386, 2013.

[124] A. Rangarajan, C. Talora, R. Okuyama, M. Nicolas, C. Mammucari, H. Oh, J. C. Aster, S. Krishna, D. Metzger, P. Chambon, et al. Notch signaling is a direct determinant of keratinocyte growth arrest and entry into differentiation. *The EMBO journal*, 20(13):3427–3436, 2001.

[125] F. Rehfeldt, A. E. Brown, M. Raab, S. Cai, A. L. Zajac, A. Zemel, and D. E. Discher. Hyaluronic acid matrices show matrix stiffness in 2d and 3d dictates cytoskeletal order and myosin-ii phosphorylation within stem cells. *Integrative biology*, 4(4):422–430, 2012.

[126] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[127] I. Rogov, I. Volkova, K. Kuleshov, and I. Savchenkova. in vitro myogenic differentiation of bovine multipotent mesenchymal stem cells taken from bone marrow and adipose tissue. , (6 (eng)), 2012.

[128] A. Saadatpour, S. Lai, G. Guo, and G.-C. Yuan. Single-cell analysis in cancer genomics. *Trends in Genetics*, 31(10):576–586, 2015.

[129] A.-E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014.

[130] M. Sasai and P. G. Wolynes. Stochastic gene expression as a many-body problem. *Proceedings of the National Academy of Sciences*, 100(5):2374–2379, 2003.

[131] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–540, 2008.

[132] Y. Shi, Z. Nenadic, and X. Xu. Novel use of matched filtering for synaptic event detection and extraction. *PLoS One*, 5(11):e15517, 2010.

[133] G. Silberberg and H. Markram. Disynaptic inhibition between neocortical pyramidal cells mediated by martinotti cells. *Neuron*, 53(5):735–746, 2007.

[134] P. Smolen, D. A. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networksmethods, recent results, and future directions. *Bulletin of mathematical biology*, 62(2):247–292, 2000.

[135] O. Stegle, L. Parts, R. Durbin, and J. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, 6(5):e1000770, 2010.

[136] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

[137] O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[138] A. J. Stops, K. Heraty, M. Browne, F. J. O'Brien, and P. McHugh. A prediction of cell differentiation and proliferation within a collagen–glycosaminoglycan scaffold subjected to mechanical strain and perfusive fluid flow. *Journal of biomechanics*, 43(4):618–626, 2010.

[139] C. H. Streuli, C. Schmidhauser, M. Kobrin, M. J. Bissell, and R. Derynck. Extracellular matrix regulates expression of the tgf-beta 1 gene. *The Journal of cell biology*, 120(1):253–260, 1993.

[140] M. Sun, F. Spill, and M. H. Zaman. A computational model of yap/taz mechanosensing. *Biophysical journal*, 110(11):2540–2550, 2016.

[141] Y. Sun, C. S. Chen, and J. Fu. Forcing stem cells to behave: a biophysical perspective of the cellular microenvironment. *Annual review of biophysics*, 41:519–542, 2012.

[142] J. Swift, I. L. Ivanovska, A. Buxboim, T. Harada, P. D. P. Dingal, J. Pinter, J. D. Pajerowski, K. R. Spinler, J.-W. Shin, M. Tewari, et al. Nuclear lamin-a scales with tissue stiffness and enhances matrix-directed differentiation. *Science*, 341(6149):1240104, 2013.

[143] Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara, and S. Miyano. Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection. *Bioinformatics*, 19(suppl 2):ii227–ii236, 2003.

[144] K. Tan, J. Tegner, and T. Ravasi. Integrated approaches to uncovering transcription regulatory networks in mammalian cells. *Genomics*, 91(3):219–231, 2008.

[145] A. M. Thomson and C. Lamy. Functional maps of neocortical local circuitry. *Frontiers in neuroscience*, 1:2, 2007.

[146] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

[147] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

[148] B. Trappmann, J. E. Gautrot, J. T. Connelly, D. G. Strange, Y. Li, M. L. Oyen, M. A. C. Stuart, H. Boehm, B. Li, V. Vogel, et al. Extracellular-matrix tethering regulates stem-cell fate. *Nature materials*, 11(7):642–649, 2012.

[149] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500):371–375, 2014.

[150] K. Tsioris, A. J. Torres, T. B. Douce, and J. C. Love. A new toolbox for assessing single cells. *Annual review of chemical and biomolecular engineering*, 5:455, 2014.

[151] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[152] N. G. Van Kampen and W. P. Reinhardt. Stochastic processes in physics and chemistry, 1983.

[153] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, et al. Self-organizing map in matlab: the som toolbox. In *Proceedings of the Matlab DSP conference*, volume 99, pages 16–17, 1999.

[154] J. Wang, C. Li, and E. Wang. Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network. *Proceedings of the National Academy of Sciences*, 107(18):8195–8200, 2010.

[155] J. Wang, L. Xu, and E. Wang. Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation, and coherence of biochemical oscillations. *Proceedings of the National Academy of Sciences*, 105(34):12271–12276, 2008.

[156] J. Wang, K. Zhang, L. Xu, and E. Wang. Quantifying the waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, 108(20):8257–8262, 2011.

[157] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, and L. Chen. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, 22(19):2413–2420, 2006.

[158] F. M. Watt, C. L. Celso, and V. Silva-Vargas. Epidermal stem cells: an update. *Current opinion in genetics & development*, 16(5):518–524, 2006.

[159] J. Wells, B. Lee, A. Q. Cai, A. Karapetyan, W.-J. Lee, E. Rugg, S. Sinha, Q. Nie, and X. Dai. Ovol2 suppresses cell cycling and terminal differentiation of keratinocytes by directly repressing c-myc and notch1. *Journal of Biological Chemistry*, 284(42):29125–29135, 2009.

[160] J. H. Wen, L. G. Vincent, A. Fuhrmann, Y. S. Choi, K. C. Hribar, H. Taylor-Weiner, S. Chen, and A. J. Engler. Interplay of matrix stiffness and protein tethering in stem cell differentiation. *Nature materials*, 13(10):979–987, 2014.

[161] J. L. Wilson, S. Suri, A. Singh, C. A. Rivet, H. Lu, and T. C. McDevitt. Single-cell analysis of embryoid body heterogeneity using microfluidic trapping array. *Biomedical microdevices*, 16(1):79–90, 2014.

[162] X. Xu and E. M. Callaway. Laminar specificity of functional input to distinct types of inhibitory cortical neurons. *Journal of Neuroscience*, 29(1):70–85, 2009.

[163] X. Xu, N. D. Olivas, T. Ikrar, T. Peng, T. C. Holmes, Q. Nie, and Y. Shi. Primary visual cortex shows laminar specific and balanced circuit organization of excitatory and inhibitory synaptic connectivity. *The Journal of physiology*, 2016.

[164] X. Xu, N. D. Olivas, R. Levi, T. Ikrar, and Z. Nenadic. High precision and fast functional mapping of cortical circuitry through a novel combination of voltage sensitive dye imaging and laser scanning photostimulation. *Journal of neurophysiology*, 103(4):2301–2312, 2010.

[165] Z. Xue, K. Huang, C. Cai, L. Cai, C.-y. Jiang, Y. Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, et al. Genetic programs in human and mouse early embryos revealed by single-cell rna [thinsp] sequencing. *Nature*, 500(7464):593–597, 2013.

[166] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139, 2013.

[167] C. Yang, M. W. Tibbitt, L. Basta, and K. S. Anseth. Mechanical memory and dosing influence stem cell fate. *Nature materials*, 13(6):645–652, 2014.

[168] Y. Yoshimura, J. L. Dantzker, and E. M. Callaway. Excitatory cortical neurons form fine-scale functional networks. *Nature*, 433(7028):868–873, 2005.

[169] G. Yourek, M. A. Hussain, and J. J. Mao. Cytoskeletal changes of mesenchymal stem cells during differentiation. *ASAIO journal (American Society for Artificial Internal Organs: 1992)*, 53(2):219, 2007.

[170] A. Zarrinpar and E. M. Callaway. Local connections to specific types of layer 6 neurons in the rat visual cortex. *Journal of neurophysiology*, 95(3):1751–1761, 2006.

[171] K. Zhang, M. Sasai, and J. Wang. Eddy current and coupled landscapes for nonadiabatic and nonequilibrium complex system dynamics. *Proceedings of the National Academy of Sciences*, 110(37):14930–14935, 2013.

[172] J. X. Zhou and S. Huang. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends in Genetics*, 27(2):55–62, 2011.

Appendices

A Additional file for Chapter 1

Regarding the relationship between SAA and S, we observe that the several orders of magnitude of stiffness range, and a hyperbolic relationship, consistent with Figs. 2C and 4B of a previous work Rehfeldt et al [125].

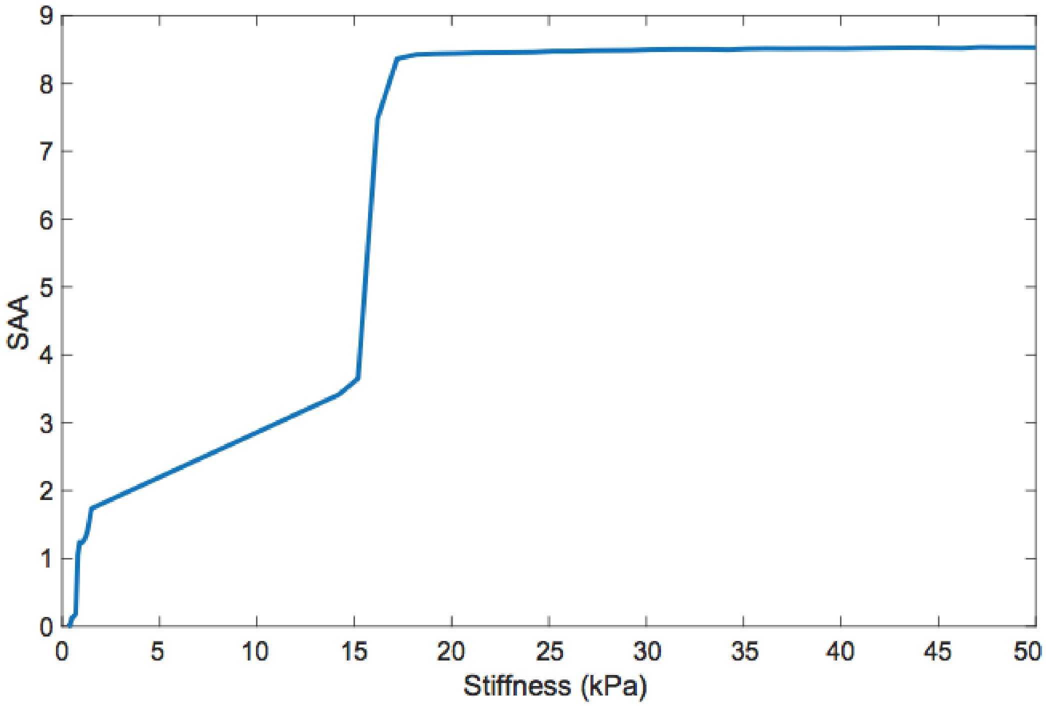


Figure A.1: The trajectory of SAA against the values of stiffness S.

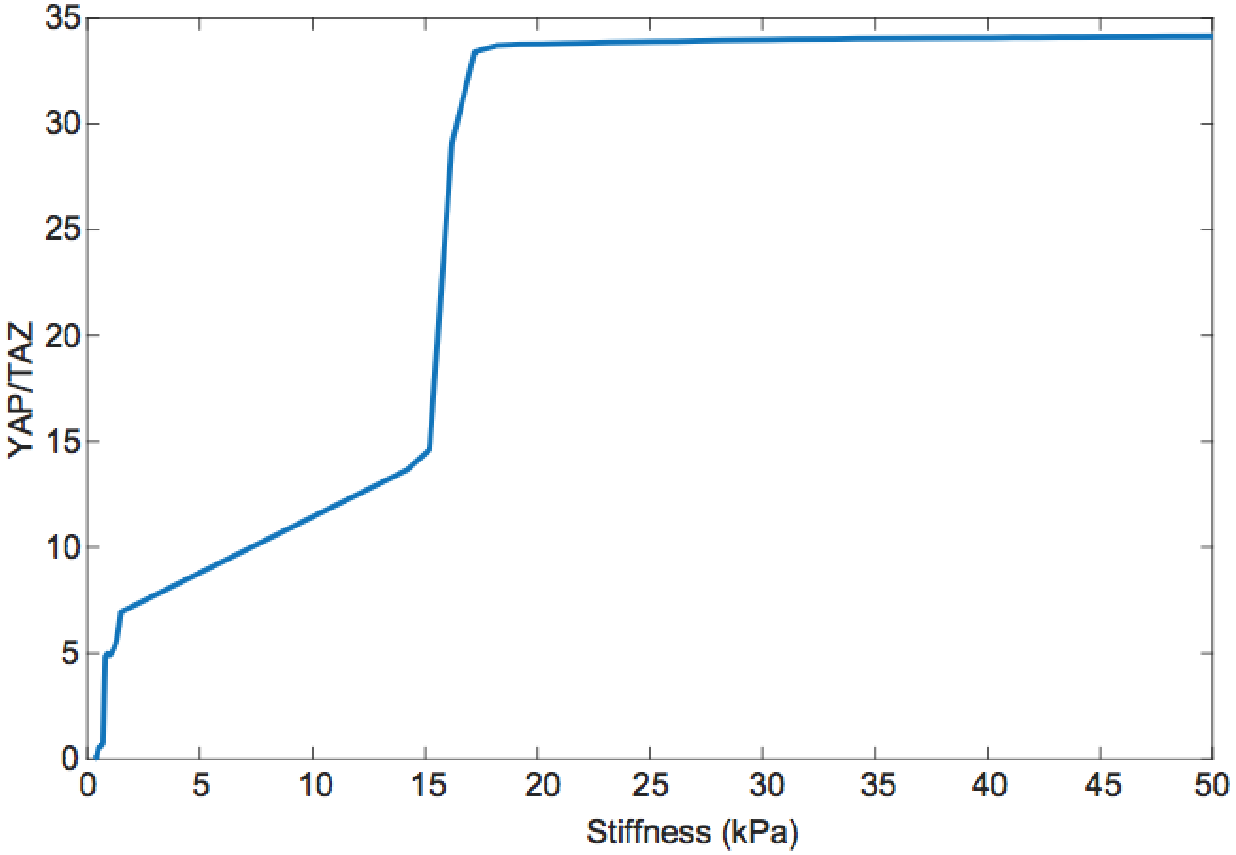


Figure A.2: The trajectory of YAP/TAZ against the values of stiffness S.

In our model, we only implicitly consider the relocalization of YAP/TAZ since the species YAP/TAZ in our model is best described as functional YAP/TAZ, i.e. the ratio of (nuclear YAP/TAZ: cytoplasmic YAP/TAZ). This is due to several observations: 1) mechano-sensing is tightly coupled to YAP/TAZ relocalization [49, 142]; and 2) nuclear YAP/TAZ is the effector form of this species given that we model its ability to modify the transcription of target genes. Below we plot the values of functional YAP/TAZ against the stiffness. The plot demonstrates a more complex relationship that that shown in Swift et al Fig. 4I [142], however again here we note the considerable differences in stiffness scales between the two works.

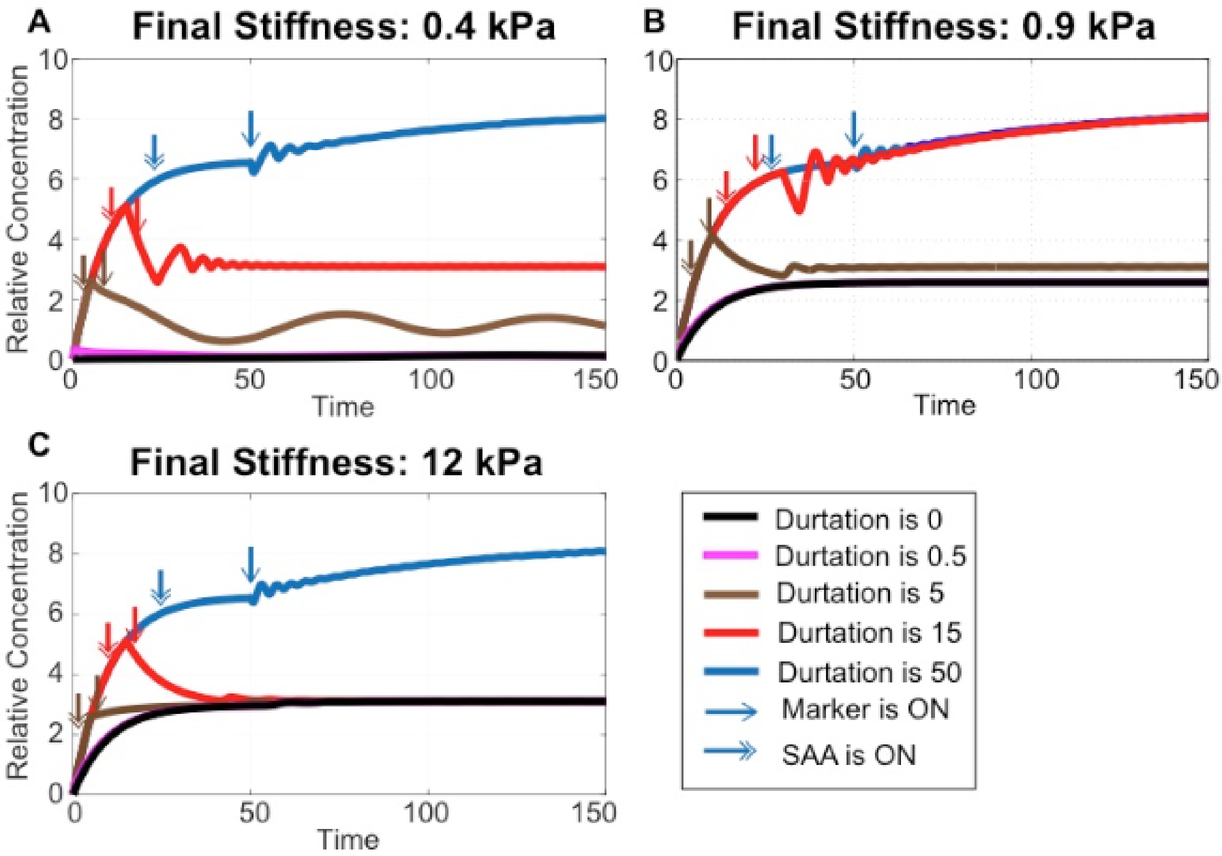


Figure A.3: Comparison of the time point when the marker genes go ON and the time point when the SAA increases significantly. The first seeding stiffness in this figure is 30 kPa. The second seeding stiffness is 0.4 kPa (A), 0.9 kPa (B) or 12 kPa (C). Here we use the single-headed arrows to illustrate when the marker genes go on and the double-headed arrow to illustrate the time point when the SAA increases significantly. The different colors are the different durations of the first seeding.

Above we have plotted the trajectory of SAA overtime under the same conditions as in Figure 1.5 in the main text. Here we use the single-headed arrows to illustrate when the marker genes go on and the double-headed arrow to illustrate the time point when the SAA increases significantly. As shown, we can see that the double-headed arrows are in each case before the corresponding single headed ones. This shows that SAA increase before differentiation, which is consistent with the observations in the previous experiments [125].

B Additional file for Chapter 2

B.1 Computing contours of CSM

In the process of identifying basins of a CSM we use contour lines to approach W_m. The total number of contour lines, N_c, determines the height difference between adjacent contour lines by $h = (max(M_{cs}) - min(M_{cs}))/N_c$. When h is smaller, it is better for contour lines to capture the topology of W_m. When h is greater, the contour lines will cross some W_m and it leads to an increase of the size of captured basins and to a decrease of the number of basins in the CSM. Figure S3 shows that the simulation results under different N_c with same values of other parameters. This parameter needs to be tuned to obtain a consistent CSM for data. In practice different datasets may need different values of N_c. The values of N_c in the different simulations are as follows. $N_c = 400$ in Figure 2 in the main text. $N_c = 23, 50, 38, 400$ in Figure 2.3A, 2.3B, 2.3C, and 2.3D in the main text, respectively. $N_c = 25, 40, 60$ in Figure 2.4, Figure 2.5, and Figure 2.6 in the main text, respectively. In addition, the values of N_c generating the figures in the Supplementary file are listed in its caption of each figure.

B.2 Stochastic differential equations model to generate the simulation data

We constructed a toy system consisting of a small number of genes to mimic the single cell gene expression data, and more effectively to evaluate the performance and choices of parameters of our algorithm. The toy system contains three stages, and in each stage one type of cells makes a transition to two other types in the next stage (Figure 2.2). Together, seven types of cells with three branches are then produced from the model. The cellular types are defined by the specific patterns of gene expression levels of the six genes that interlinked (Figure 2.2A). Specifically, in Type 1 cells Gene A and Gene B are activated and all other four genes are silenced; In Type 2 cells, Gene A, Gene C and Gene D are active, and in Type 3 cells Gene B, Gene E and Gene F are activated; When one of Gene A and Gene B and one of Gene C, Gene D, Gene E and Gene F are active, four different other types of cells in the third stage are defined respectively.

The three-toggle modules of the six genes are then modeled using stochastic differential equations with noise introduced into the system [19].

$$dX(t) = F(X(t), \Theta) + \eta dW(t) \tag{B.1}$$

X(t) is a vector $(x_A(t), x_B(t), x_C(t), x_D(t), x_E(t), x_F(t))^T$, $x_*(t)$ is the expression level of gene $*$ at time t. η is the variance vector and $W(t) = (w_1(t), w_2(t), w_3(t), w_4(t), w_5(t), w_6(t))^T$ is the scalar white noise (Wiener process), $F(X(t), \Theta) = (f_i(X(t), \Theta))_i$ where $i = 1, 2, ..., 6$.

$$f_i(X(t) = \alpha_i \prod_{j=1}^{6} g_j(x_j(t).\Theta) - \beta x_i(t) \tag{B.2}$$

$$g_j(x_j(t).\Theta) = \begin{cases} \frac{x_j(t)^n}{x_j(t)^n + k^n} & \text{if } x_j \text{ is an activator} \\ \frac{k^n}{x_j(t)^n + k^n} & \text{if } x_j \text{ is an inhibitor} \end{cases} \tag{B.3}$$

Where n and k are the entries of the parameter vector Θ. The standard Euler-Maruyama method is employed to solve the stochastic differential equations (Eq. 1.19) [54].

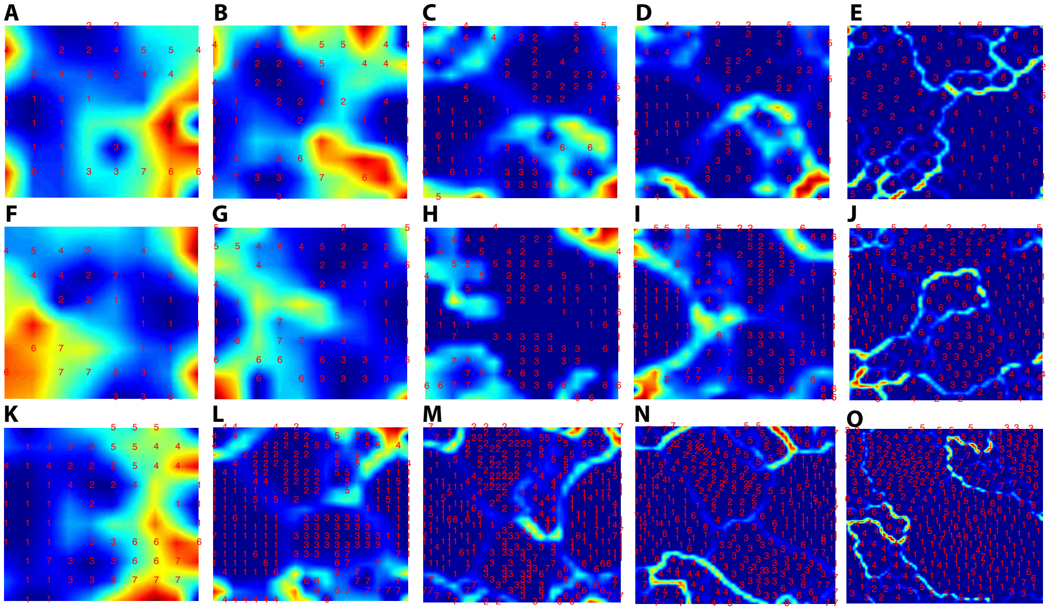


Figure B.4: The CSMs are calculated using different numbers of simulated observations and different sizes of maps in the SOMSC. From A to E, $N_g = 64, 100, 225, 400, 2500$ with the same $N = 100$ observations, respectively. From F to J, $N_g = 64, 100, 225, 400, 2500$ but with the same $N = 200$ observations, respectively. From F to J, $N_g = 100, 400, 1089, 2500, 6400$ using the same $N = 353$ observations, respectively. Here $U_0 = 1.5$, $N_c = 400$ and $\gamma = 1$ for simulations.

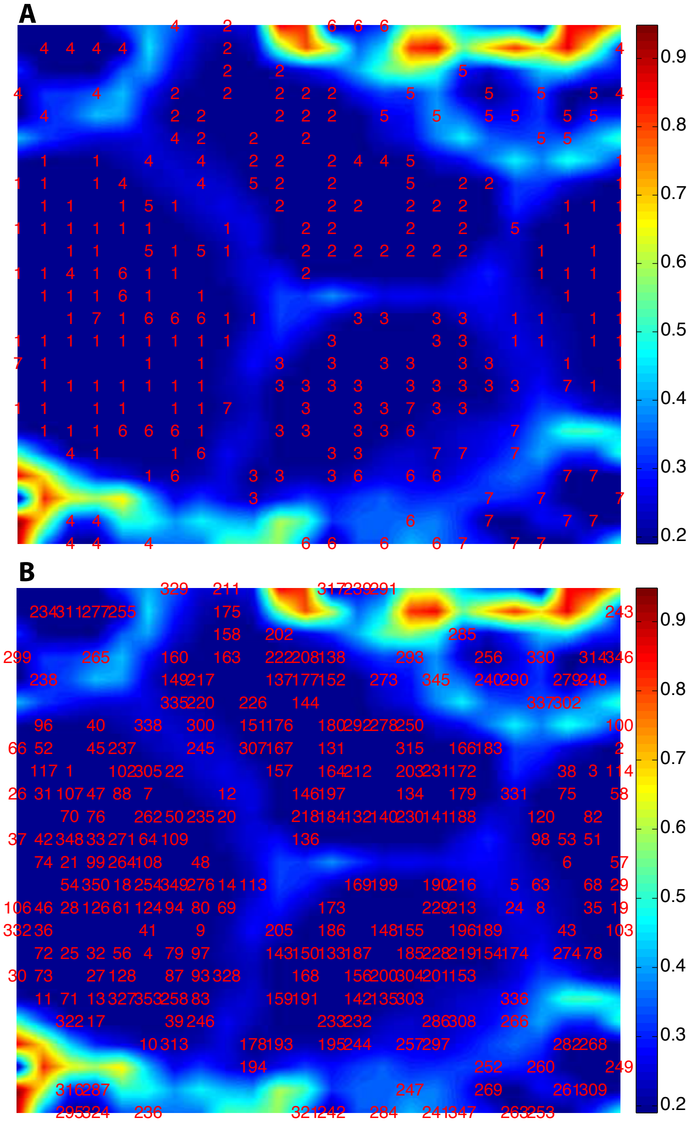


Figure B.5: The CSM of simulation data with $N = 353$ cells. Here $N_g = 576$, $N_c = 400$, $U_0 = 1.5$, and $\gamma = 1$. (A) The CSM showing the distribution of cell stages. A red number is a temporal stage of its corresponding cell at that point. (B) The CSM showing the distribution of cells with cell index. A red number is an index of its corresponding cell at that point. For example, '205' means that the 205th cell is located at that position.

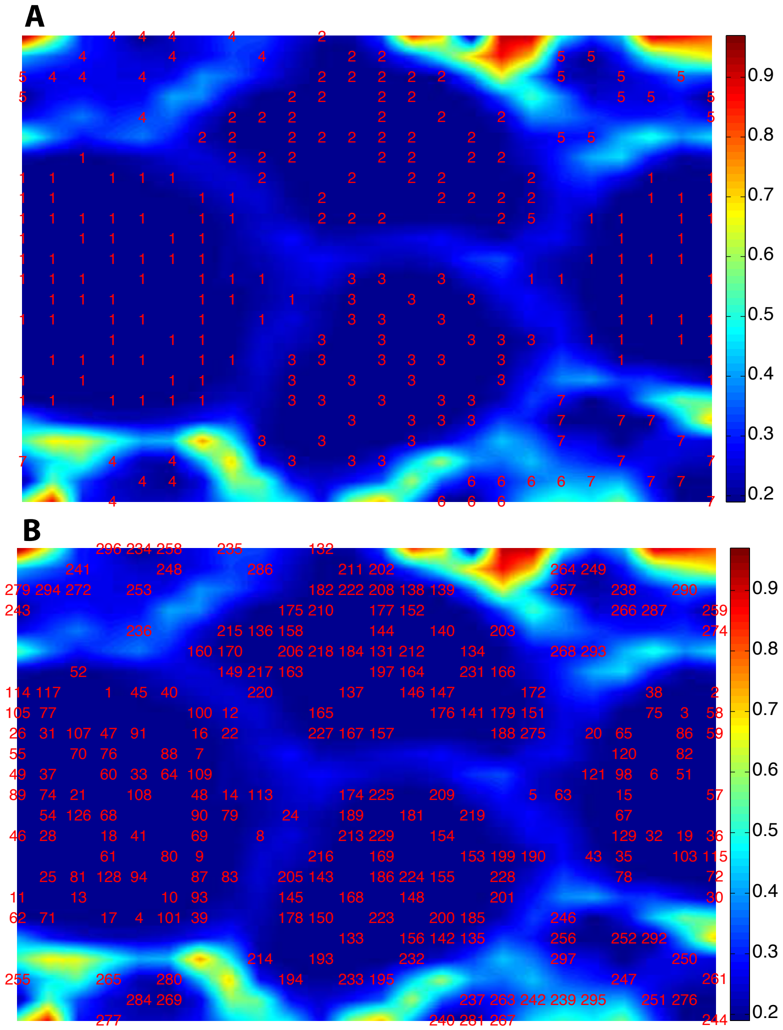


Figure B.6: The CSM of simulation data with $N = 279$ cells. Here $N_g = 576$, $N_c = 450$, $U_0 = 1.5$, and $\gamma = 1$ after removing the cells located in incorrect basins from the original data of $N = 353$ cells. (A) The CSM showing the distribution of cell stages. A red number is a stage of its corresponding cell at that point. (B) The CSM showing the distribution of cell index. A red number is an index of its corresponding cell at that point.

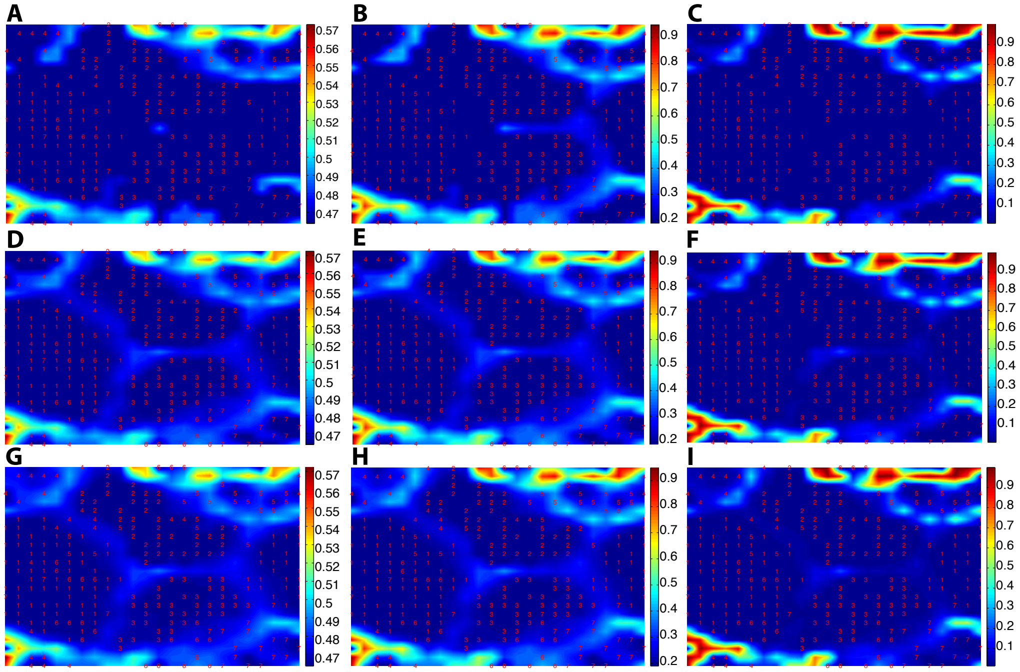


Figure B.7: The CSMs of simulation data with $N = 353$ and $U_0 = 1.5$ under different γ and N_c. A red number is a temporal stage of its corresponding cell at that point. (A) $\gamma = 0.1$ and $N_c = 100$. (B) $\gamma = 1$ and $N_c = 100$. (D) $\gamma = 3$ and $N_c = 100$. (D) $\gamma = 0.1$ and $N_c = 450$. (E) $\gamma = 1$ and $N_c = 450$. (F) $\gamma = 3$ and $N_c = 450$. (G) $\gamma = 0.1$ and $N_c = 2000$. (H) $\gamma = 1$ and $N_c = 2000$. (I) $\gamma = 3$ and $N_c = 2000$.

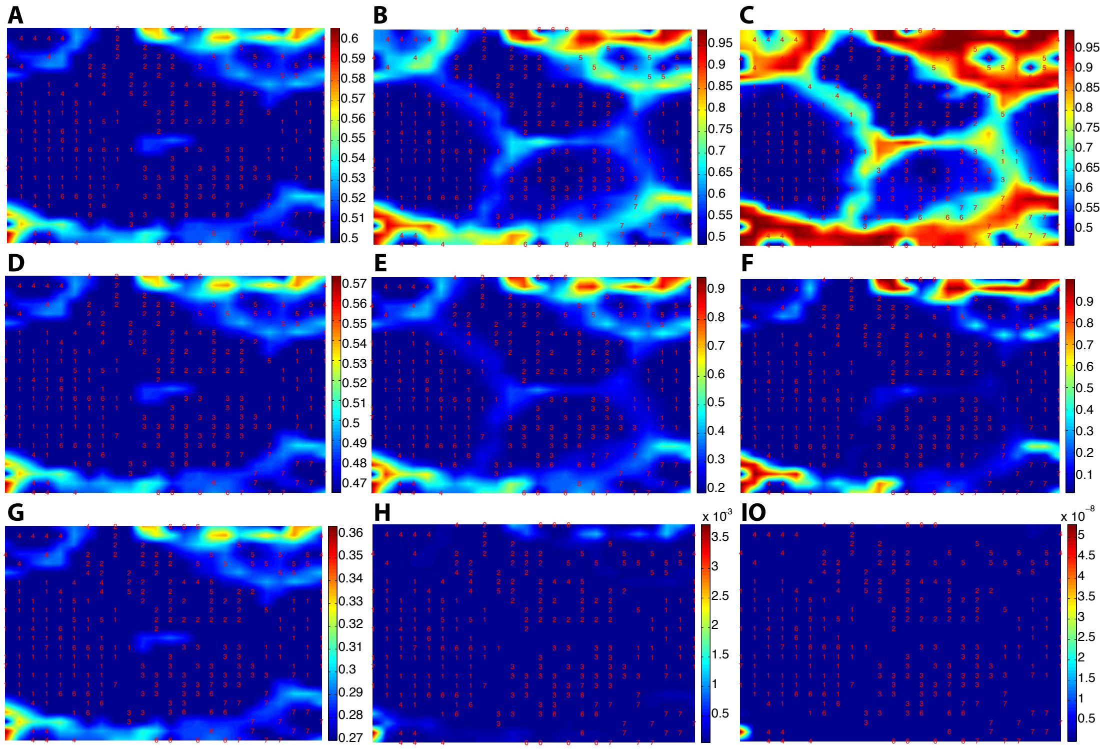


Figure B.8: The CSMs of simulation data with $N = 353$ and $N_c = 400$ under different γ and U_0. A red number is a temporal stage of its corresponding cell at that point. (A) $\gamma = 0.1$ and $U_0 = 0.1$. (B) $\gamma = 1$ and $U_0 = 0.1$. (C) $\gamma = 3$ and $U_0 = 0.1$. (D) $\gamma = 0.1$ and $U_0 = 1.5$. (E) $\gamma = 1$ and $U_0 = 1.5$. (F) $\gamma = 3$ and $U_0 = 1.5$. (G) $\gamma = 0.1$ and $U_0 = 10$. (H) $\gamma = 1$ and $U_0 = 10$. (I) $\gamma = 3$ and $U_0 = 10$.

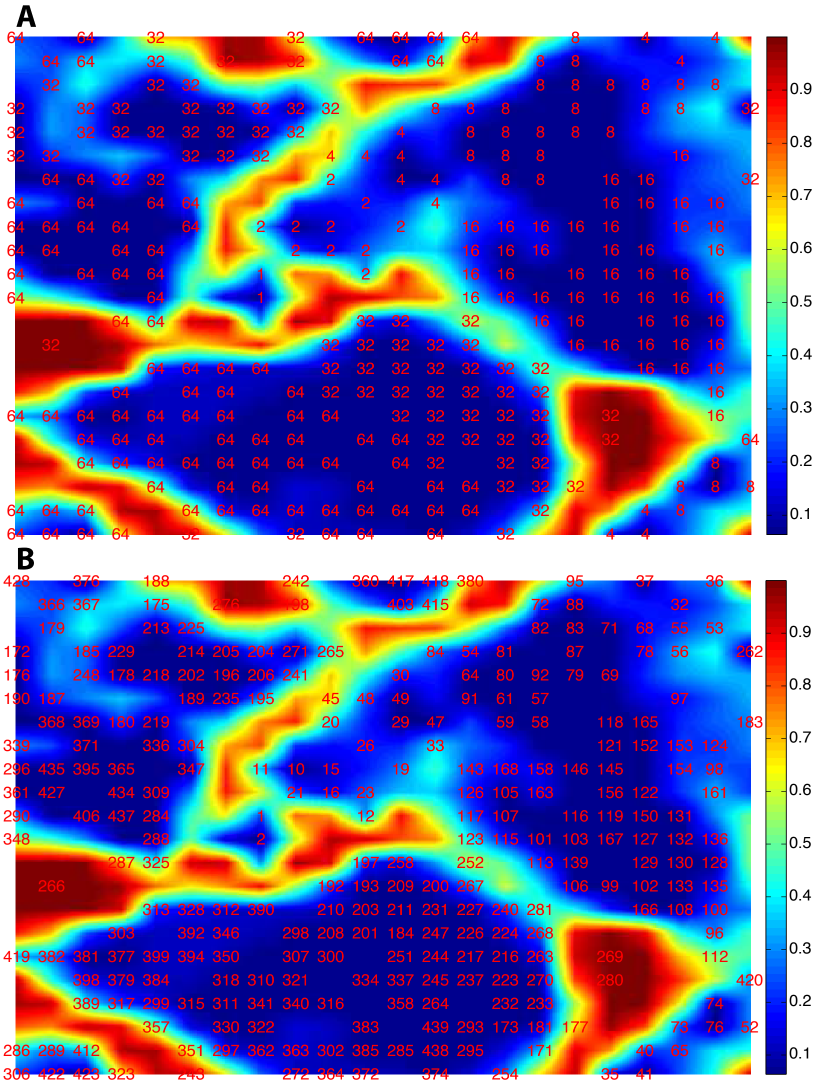


Figure B.9: The CSM of the data from mouse stem cells from zygote to blastocyst with $N = 442$. Here, $N_g = 484$, $N_c = 400$, $U_0 = 2$, and $\gamma = 2$. (A) The CSM showing the distribution of cell types. A red number is a temporal stage of its corresponding cell at that point. (B) The CSM showing the distribution of cells with cell index with the given numbers in the measurements. A red number is an index of its corresponding cell at that point.

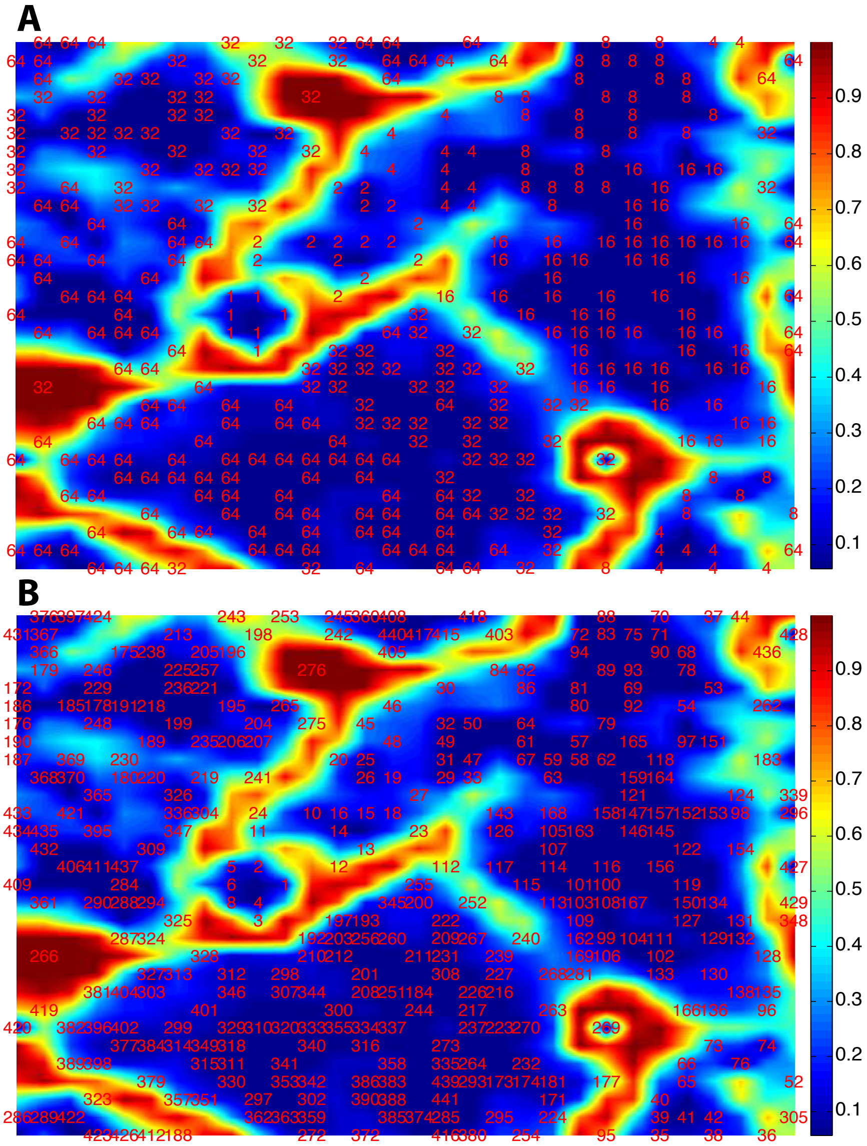


Figure B.10: The CSM of the data from mouse stem cells from zygote to blastocyst with $N = 442$. Here, $N_g = 900$, $N_c = 400$, $U_0 = 2$, and $\gamma = 2$. (A) The CSM showing the distribution of cell stages. A red number is a temporal stage of its corresponding cell at that point. (B) The CSM showing the distribution of cells with cell index. A red number is an index of its corresponding cell at that point.

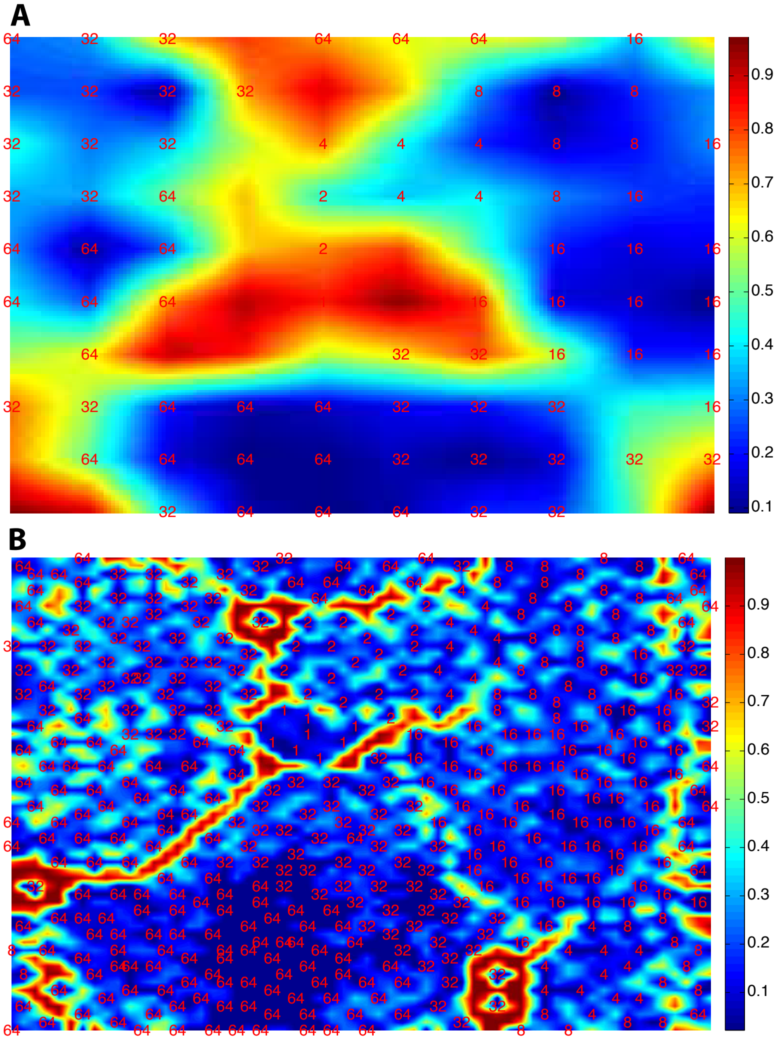


Figure B.11: The CSM of the data from mouse stem cells from zygote to blastocyst with $N = 442$ under different N_g. (A) The CSM showing the distribution of cell stages with $N_g = 100$, $N_c = 400$, $U_0 = 2$, and $\gamma = 2$. (B) The CSM showing the distribution of cell types with $N_g = 3600$, $N_c = 400$, $U_0 = 2$, and $\gamma = 2$. A red number is a temporal stage of its corresponding cell at that point.

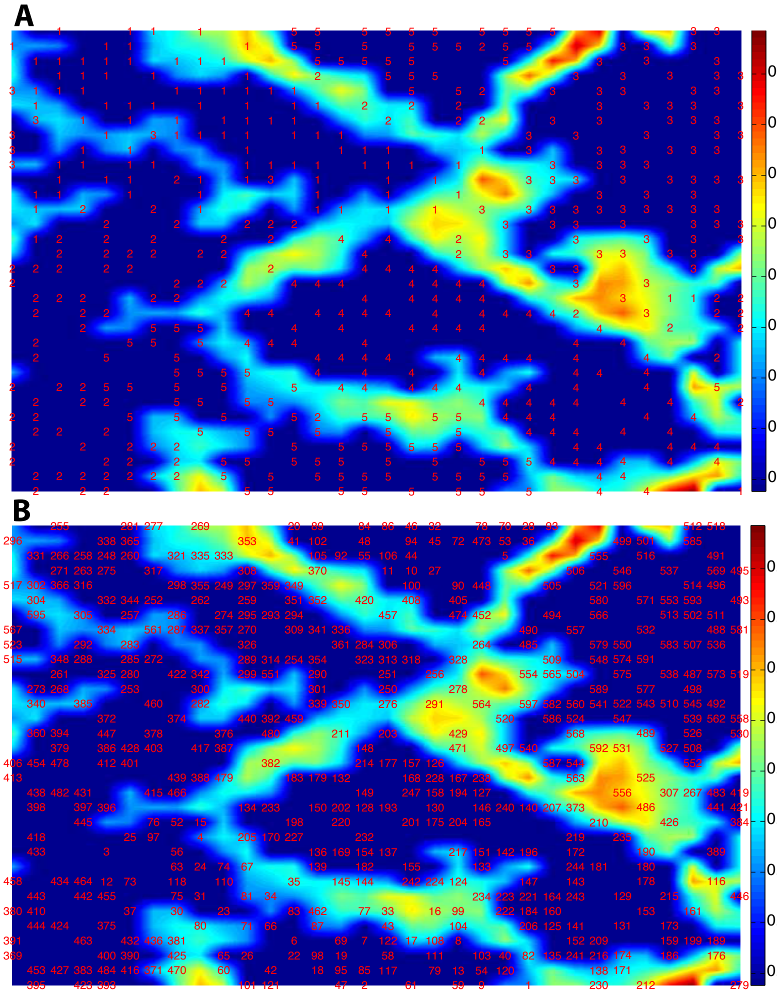


Figure B.12: The CSM of the data from mouse haematopoietic stem cells with $N = 597$. Here $N_g = 1024$, $N_c = 25$, $U_0 = 1.5$, and $\gamma = 0.88$. (A) The CSM showing the distribution of cell types. A red number is a temporal stage of its corresponding cell at that point . (B) The CSM showing the distribution of cells with cell index. A red number is an index of its corresponding cell at that point.

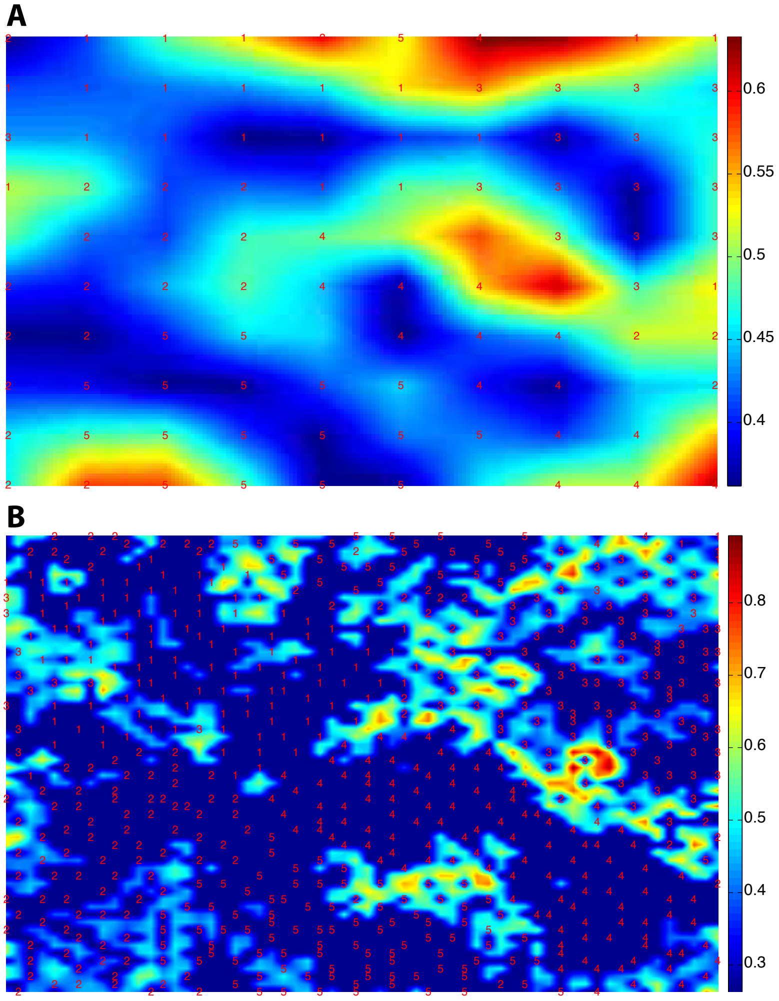


Figure B.13: The CSM of the data from mouse haematopoietic stem cells with $N = 597$ under different N_g. (A) The CSM showing the distribution of cell types as $N_g = 100$, $N_c = 250$, $U_0 = 1.5$, and $\gamma = 0.88$. (B) The CSM showing the distribution of cell types as $N_g = 3600$, $N_c = 250$, $U_0 = 1.5$, and $\gamma = 0.88$. A red number is a temporal stage of its corresponding cell at that point.

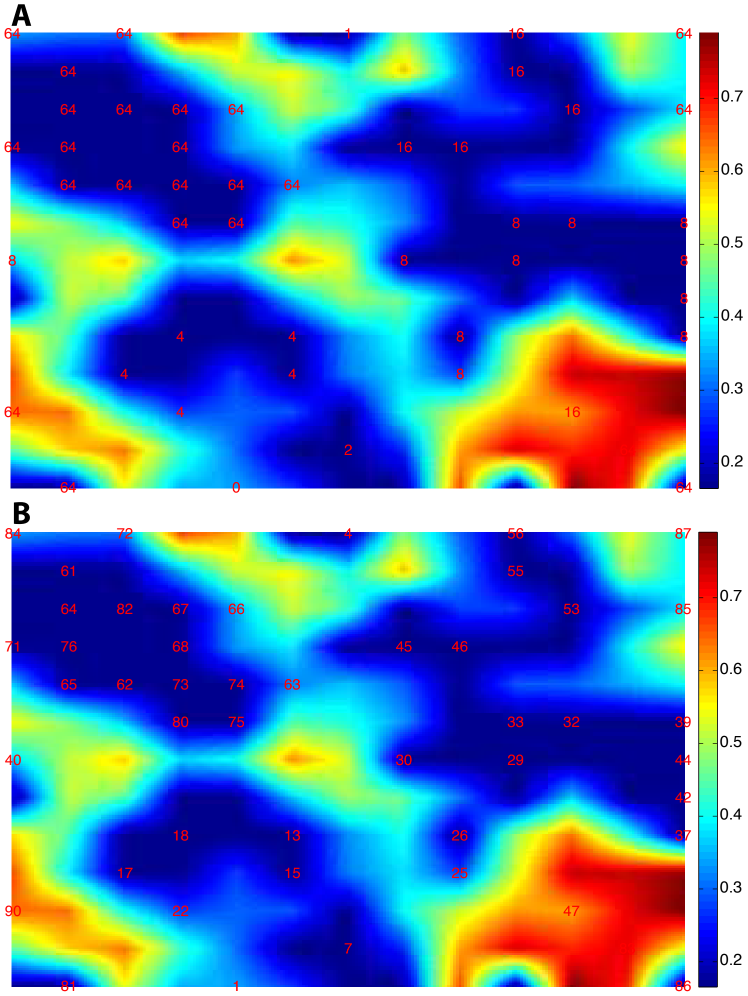


Figure B.14: The CSM of the data from human preimplantation embryonic cells from oocyte to late blastocyst with $N = 90$ as $N_g = 169$, $N_c = 40$, $U_0 = 20$, and $\gamma = 0.1$. (A) The CSM showing the distribution of cell stages. A red number is a temporal stage of its corresponding cell at that point . (B) The CSM showing the distribution of cells with cell index. A red number is an index of its corresponding cell at that point.

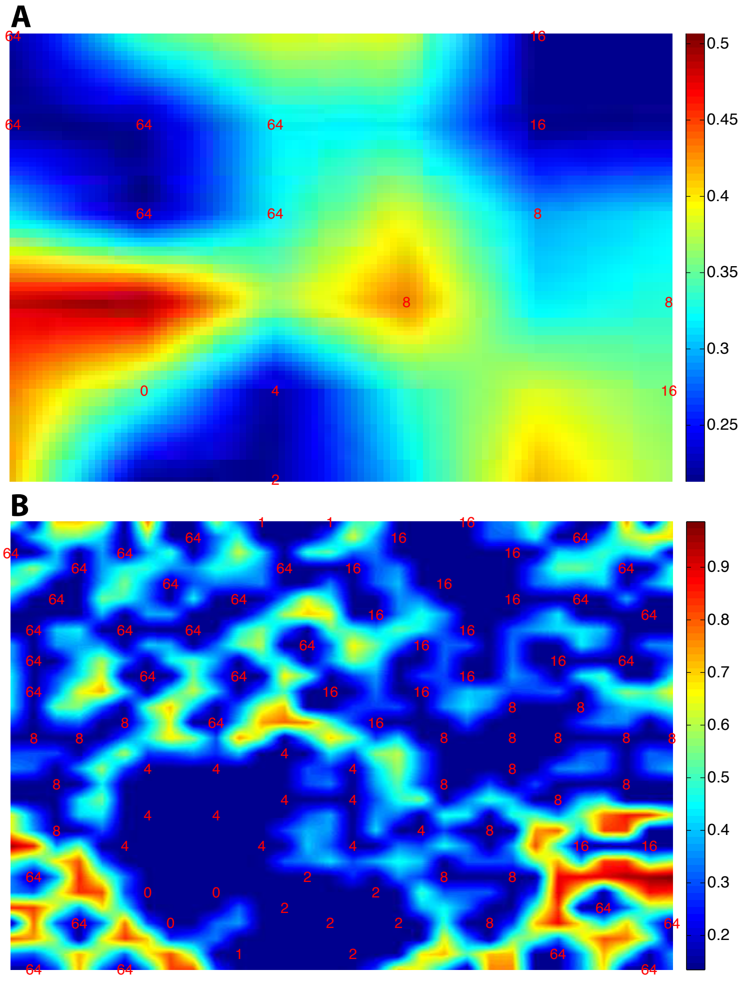


Figure B.15: The CSM of the data from human preimplantation embryonic cells from oocyte to late blastocyst with $N = 90$ under different N_g. (A) The CSM showing the distribution of cell types as $N_g = 36$, $N_c = 40$, $U_0 = 20$, and $\gamma = 0.1$. A red number is a temporal stage of its corresponding cell at that point. (B) The CSM showing the distribution of cell types as $N_g = 900$, $N_c = 40$, $U_0 = 20$, and $\gamma = 0.1$. A red number is an index of its corresponding cell at that point.